統計如何說謊

[1]

自學書院　譯著

2014



[1] http://www.surgerycenterrecruiter.com/wp-content/uploads/2012/08/Leukemia-Statistics-lg.jpg

譯言

看了第一本有關統計謊言的著作 *How to Lie with Statistics by Darrell Huff, 1954*，立論精闢，雖然書中一些例子已經過時，理據依然對照現在的「統計誤世」年代。電腦軟件又引進了一些新工具和誤區。考慮之下，為保留原作面貌，選擇譯本每章分為兩部份。第一部份翻譯原書（略有刪節，省掉沒有歷史背景資料很難明白的例子），第二部份選譯補充材料，主要參考如何利用統計數據撒謊（WikiHow）、統計學〈維基百科〉、統計誤用〈維基百科〉、Misleading graph〈維基百科〉以及其他網頁資料。

譯本以 Creative Commons 條款發表，即是：保留署名權(Attribution)，歡迎各位下載、轉載和分發，允許衍生作品（必須以相同條款分發 Share Alike）和禁止商業用途(Non-commercial)條款發表。

Creative Commons 有限版權制度面世已經十年，全球有一百三十多個國家和地區已有本土化的 Creative Commons 條款。Creative Commons 條款適用於任何創作成果：大如維基百科，YouTube 視頻、Flickrs 相片集，小如個人網誌，都可以是以 Creative Commons 條款發表的學習和應用材料。如各路英雄一呼百應，本著知識共享的精神，壯大 Creative Commons 的範疇，互相支持，互補互助，網上的知識源泉定必波瀾壯闊。

華文世界的 Creative Commons 發展，有是有，但比諸其他語言，實在落後於人。「革命尚未成功，同志還需努力。」

關於「統計學」的 Creative Commons 著作，我只找到劉彥方和陳強立的《思方網 ： 統計與圖表》，如高人有其他發現，請告知。

自學書院
2014 年 6 月

統計的重要



複雜的現代社會離不開調查和統計。相關人員收集、整理、歸納、分析數據和發表結果，廣泛應用在自然科學、社會科學和人文科學，也用於決定工商業及政府政策。日常生活躲也躲不了的廣告也每每以統計數據引導消費者。

統計是為面對不定狀況制定決策提供方法的科學。統計學和機率論的關係異常密切，事實上任何統計問題的研究都必涉及機率論的運用，後者實為前者的主要工具。統計可以是利用現有數據或通過調查取得數據。除非**母體群**[2](population)規模特小，調查可以覆蓋全部，一般調查是以取樣方式進行：搜集小量數據（樣本 sample）的資料以估計、預測和研究母體群。

統計陷阱帶來的負面影響可大可小。基於錯誤統計的政策可能差之毫厘，謬以千里；醫學的統計陷阱可能要數十年後才被糾正，招致人命損失。近代廣告特多統計數字引導誤導消費者。

要了解統計的諸多陷阱，先看看一般統計的流程。

利用現有數據的統計主要是案頭作業，這方面的陷阱亦見諸調查統計。要搜尋未知的數據，抽樣調查是最常用的搜集方法。

一般而言，統計作業的步驟如下：
1. 決定調查主題。
2. 決定收集資料的方法：(a)書面作業或(b)調查：面對面訪問，郵寄問卷、電話訪問或混合運用。
3. 界定(a)書面作業的範圍或(b)抽樣調查的母體群。
4. 決定(b)抽樣使用的母體群清冊：如電話號碼簿、會員名單、戶籍資料等。
5. 決定(b)抽樣方式：隨機抽樣、分層抽樣、系統抽樣或分段抽樣。

[2] 亦作 parent population, universe；有譯為「總體、母體、母群 」。

6. 決定(b)樣本大小；若需分層，需決定分層方式及各層樣本大小。
7. (b)進行抽樣，選取樣本元素。
8. 設計(b)收集資料的形式；設計調查問卷，預試。
9. (a)彙集資料；(b)執行調查，向樣本收集反饋。
10. (a)和(b)資料檢誤、處理及分析。
11. (a)和(b)發表結果。

從上可見，每一步驟都涉及人為因素和諸多可操控手段。無論是什麼形式的統計，都可能出錯；這可能是意外，也可能是故意，構成統計陷阱。

有三種謊言：謊言，該死的謊言和統計數字。﹏Benjamin Disraeli

總有一天，有教養的公民能讀能寫，也要有統計思維。﹏H. G. Wells

我們不知道的那些事情不會讓我們陷入困境，
而是我們知道但並非如此的事情。﹏Artemus Ward

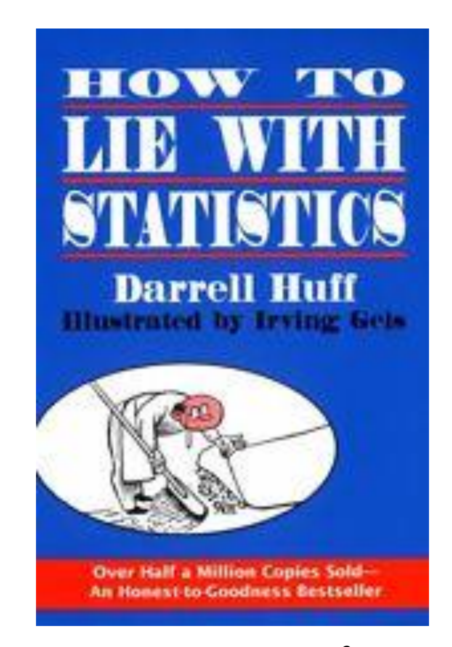
數字與統計

「多數人對於數字具有先天的畏懼感，是有演化的根源；因為人類存活在地球有幾十萬年，大多數時候是幾十人、最多百來人的小族群過著狩獵採集的生活，全部家當兩隻手就可拿著走，因此不需要用上什麼數字，對成千上萬的大數字更是沒有概念。只有在近一萬年來，人類採行農業生活後，人類社會的規模與財富不斷累積成長，才開始出現對數字的需求，也才有天賦異稟之士發展出各式各樣的數學。

雖然多數人對數字可能無感，但冰冷的數字還是要比感性的言語可靠。統計是整理大數字的科學方法，如果是因為不懂統計，或吃過統計的虧，就把統計與謊言並列，可說是因噎廢食，也算另一種人的偏見吧。」
﹏〈潘震澤：人類天生的缺陷：數字盲〉

引文說「把統計與謊言並列」是「另一種人的偏見」。相信沒有人會把全部統計看作為謊言，但統計有誤區，也不能否認有人利用統計說謊。統計有什麼誤區？如何說謊？這是本書的主題。

統計如何說謊？



Darrell Huff, 1954[3]

目錄

[3] 原文：How to Lie with Statistics by Darrell Huff, 1954。譯本略有刪節，減掉一些不懂歷史背景很難明白的過時例子。

序言

神聖古老的英國度量衡制度快要取消，英寸和英尺的時代快要結束；蓋洛普民意以一貫方式測試人們對取而代之的公制的認識，發現大學程度的男女有 33 %從未聽過公制。

然後一份周刊的讀者調查宣布讀者有 98 %知道公制。對此，報刊吹噓它的讀者群比一般人「更懂行」 。

兩項民調如何能夠有這麼明顯的差異？

蓋洛普調查員精心挑選了公眾的樣本並約見會談。這家報刊兒戲和經濟地依靠讀者填寫和郵寄問券。

由此不難猜測大部分不知道公制的讀者根本沒有興趣填報和郵寄問券，自動不參加調查。用統計術語來說，這樣的自我選擇只會產生具偏見或不具代表性的樣本，多年來導致許多許多誤導性結論。

多年前的冬季，十多位獨立調查員報告抗組織胺藥片的數量，各人都發現藥片治愈大多數感冒病例。

於是廣告和醫療產品的熱潮開始炒得火熱。這是基於人們對靈丹妙藥的永恆希望，也沒有超越統計數據去看看長久以來我們已經知道的事實。幽默作家 Henry G. Felsen 不是醫學權威，很久之前已指出適當的治療可以在七天治愈感冒：只要多休息，置諸不理，一星期就會好轉。

因此，你讀到的和聽到的平均值、關係、趨勢和圖表並不是表面的真實無誤，背後可能有更多或更少的訊息。

在追求事實的文化中，統計的祕密語言是如此吸引人，實則是用來炒作，誇大，混淆和簡單化。在報告社會和經濟趨勢、企業經營狀況、「民意」調查和人口普查的大量數據，統計方法和統計術語是必要的，但報告者用辭必須誠實和易於了解，讀者也知道用辭的意思，才不會陷於語義的無稽之談。

科普文章濫用統計數字，幾乎排擠了在半明不亮實驗室日以繼夜辛勤研究的白袍英雄。統計資料粉飾許多重要的事實，猶如撲粉化妝，上油塗漆。精心包裝的統

計勝於希特勒的「大謊言」；只是誤導，但難以追究。

這本書是如何使用統計數據來欺騙的讀本，可能看起來太像騙子手冊。也許我的好理由是類比退休竊賊出版的回憶錄等同如何挑鎖和消聲毀跡的研究生課程：騙子都知道這些技巧，老實人必須為了自衛而學習。

第一章　有內置偏差的樣本

桶內有紅豆白豆，有一種辦法肯定各有多少：倒出來點數。

有一個更簡單的辦法算出有多少紅白豆。假設桶內的紅豆白豆是相同比例，拿出一把豆子，只計數這一把。就大多數目的而言，如樣本足夠大和選擇正確，這足以代表整體。但如兩方面有偏差，其準確度可能遠遠及不上聰明的猜測，只不過是所謂科學精確的虛言。樣本因為選擇的方法有失偏頗，或過小，或兩者兼而有之，會導致謊言，也就是我們讀到或以為我們知道那些很多結論背後的可悲事實。

樣本如何出現偏差？請看一個極端的例子。假設要發問卷調查，其中包括以下的問題：「你是否喜歡回答問卷調查？」之後收回的問卷極有可能得出這樣的結論：「典型的樣本人口絕大多數喜歡回答問卷調查」，其準確度可計算至幾個小數點。這是什麼一回事？當然是因為回收的問卷已排除了大多數可能回答「不喜歡」的問卷，調查問卷已掉在廢紙簍。即使原始樣本中十有八九是「不喜歡」那幫人，這些「錯誤」已排除在外。

現實生活中是否有這樣的有偏樣本？當然有。

不久前，報刊和新聞雜誌報導在過去十年有約四百萬美國耶教舊教信徒改信新教。消息來源是跨宗派《耶教導報 *Christian Herald*》編輯 Daniel A. Poling 牧師的調查。《時代》周刊總結這故事：

> 《導報》的數字來自美國新教牧師，2,219 位牧師填報（發出問卷 25,000 份），呈報共有 51,361 前舊教教徒在過去十年加入新教。Poling 依樣本估算在十年有 4,144,366 名舊教教徒改信新教。Will Oursler 主教寫道：「即使估算有出入，全國數字不會少於二，三百萬，極有可能接近五百萬。」

雖然《時代》有報導調查中超過 90 %牧師沒有填報問卷，但錯過了指出這事實的重要性，依然精神可嘉。要徹底摧毀這調查，唯一要注意的合理可能性是大多數牧師扔掉問卷是因為沒有改信教徒的數字可以呈報。

利用這假設和 Poling 採用的相同數字（181,000 名牧師），可以另行推算。他的調查涵蓋 181,000 牧師的 25,000 人，呈報 51,361 人改信新教；如調查涵蓋全部 181,000 牧師會得出有約 370,000 人改信新教。

這樣的粗糙方法得出非常可疑的數字，但至少是一如前一數字值得信任；那個全國數字是修正數字的十一倍，因此更引人注目。至於 Oursler 主教對誤差的自信，如果他發現了一種方法來糾正未知大小的誤差，將會造福統計界。

在這背景下，多年前有另一則新聞報導，當時的幣值較高：耶魯大學學生平均年收入有$25,111。很棒！

且慢。這令人印象深刻的數字是什麼意思？這是否表明如果子女進讀耶魯或牛津，劍橋，你和他不用年老時上班？

第一眼看過去，這數字有兩個疑點：令人驚訝的精確，也不大可能這樣的令人稱羨。

只有極小可能性可以精確得知任何散漫群體以往任何時候的平均收入，更不要說精確至$111。除非收入全來自薪金，很少人能如此精確知道自己的年收入。有這樣收入的人往往會分散投資。

此外，這個可愛的平均數無疑是源於耶魯畢業生的自報收入。即使耶魯大學在 1924 年校風純樸，但不能保證四分之一世紀後這些畢業生都如實自報收入。被問及他們的收入，有些人因虛榮心或樂觀誇大了。其他人少報，尤其是擔心納稅申報，不想在任何其他文件留下自相矛盾的資料。誰知道稅務局會否看到？吹噓和低估這兩種傾向可能相互抵消，但其實是不可能的。其一傾向可能遠遠強於另一，但不知道是哪一個。

先說一下：常識告訴我們這數字幾乎不是真相。這資訊表示一些人的「平均收入」是$25,111，而這些人的實際平均收入可能較接近一半。現在看看資訊可能來源的最大誤差。

常識告訴我們不可能在二十五年後與當年的全部畢業生保持聯絡。有人已往生，有人地址不詳。

那些有通訊地址的，很多人不會回答問卷，特別關乎相當個人的資料。對於某些類型的郵件問卷，5-10%的反應已是相當高的。這一個調查的回報率應該比這更好，但肯定不是 100%。

因此，這收入數字源自有已知地址而又樂意填報個人收入的畢業生。這是否具代表性的樣本？也就是說，是否可以假設這群組的收入是相等於沒有參加調查（沒有地址或不願回報）的另一群畢業生？

在耶魯名錄，那些畢業生「地址不詳」？是否那些賺大錢的華爾街巨子，公司董事，製造業及公用事業主管？不，富人的通訊地址不難查得到。即使他們忽略了聯繫校友辦公室，從名人錄和其他參考刊物找出他們的通訊地址應是輕而易舉。二十五年後失聯的畢業生，按常理猜測應是那些畢業後事業不順的畢業生：文員，技工，流浪漢，失業酗酒漢，僅堪糊口的作家和藝術家。可能幾個人的收入總和才可攀上$ 25,111 的收入水平。他們不那麼經常參加舊生聯誼活動，可能有些人甚至不能負擔旅費。

誰會把問卷攆到垃圾桶？不能肯定，但公平的猜測至少是很多人沒有掙多多的錢可以自我吹噓。這有點像新員工發現第一份工資單夾著紙條，建議他保密工資數額，不與同事交換機密資料。這傢伙會告訴老闆：「別擔心，我和你一樣為此感到羞恥。 」

看來很清楚樣本省略了最有可能壓低平均水平的兩組。那個$25,111 數字開始為自己解釋。這只適用於有已知地址，又願意公開本人收入的特殊群體。這還要假設他們是說真話的君子。

不要輕易作出這樣的假設。抽樣調查的一個品種即是所謂「市場調研」，其經驗表明根本不能作出假設。有一項市場調查的關鍵問題是：你家看什麼雜誌？結果列表和分析顯示很多人喜愛高端的 *Harper's*，這雖然不算是曲高和寡，但至少算得是中上階層口味；並沒有很多人自認是低俗雜誌 *True Story* 的讀者。然而，出版商的數字很清楚表明 *True Story* 的發行量有幾百萬份，而 *Harper's* 只有幾十萬。調查的設計人員自我解困：也許我們問錯了對象。但事實不是這樣。調查在全國各地街上訪問。那麼唯一合理的結論是很多受訪者回答這些問題時沒有說實話。調查只是發現了人們在裝腔作勢，裝模作樣。

最終發現，如果想知道某些人看什麼雜誌，查詢是沒用的。更好的辦法是從他們家裡買入舊雜誌，這中自有資訊。

只需數算《耶魯評論》和《愛情周刊》的冊數。即使這樣也不能確實知道人們在看什麼，只是知道他們接觸什麼。

同樣，讀到有報導一般人（最近聽的很多，大部份不可信）刷牙每天一到兩次（我隨意取一個數字），這有什麼問題？誰能知道這些事情？ 女生看了無數廣告，印象中以為不刷牙是社會罪行，她會否向陌生人承認她不經常刷牙？這樣的統計只意味著人們對刷牙的說法，但沒有弄清楚人們刷牙的頻率。

諺語有云：河水向下流，不高於源頭。嗯，這似乎是可能的，如果有泵站幫忙。同樣真實的是抽樣調查的結果不會優於樣本本身。數據經通過層層統計處理，過濾為小數點平均值，調查結果開始蒙上可信的光環，但仔細看看採樣就可以否定這假像。

可信的採樣報告必須採用具代表性的樣本，即是已去除每一偏見的源頭。上文的耶魯數字頓見毫無價值。許多報刊和雜誌報導犯下同樣錯誤，沒有什麼意義。

有一次，精神科醫生報告謂幾乎每個人都是神經質。這樣的說法除了破壞「神經質」一詞的任何意義，倒不如看看這位醫生的樣本，也就是說這位精神科醫生一直在觀察什麼人？原來，他是從觀察他的病人得出這啟發性結論；這個「樣本」根本不能作為總體人口的樣本。正常人不會看心理醫生的。

閱讀不要囫圇吞棗，可以避免學習了一大堆表裡不一的東西。

值得銘記無論是有形或無形來源的偏差都會破壞樣本的可靠性。也就是說，即使不能找到可證實偏見的來源，只要有偏差的可能性，對結果也應保持一定程度的懷疑。

一項例證是 1936 年《文學文摘》月刊的著名慘敗。月刊的一千萬名電話用戶和月刊訂戶調查曾準確預測 1932 年的總統大選。1936 年，月刊彙集同一名單的反饋，編輯部放心預測羅斯福只有 161 選舉人票，對手 Landon 得票 370。這樣本名單久經測試，怎會有偏差？當然有偏差；無數高校論文和其他事後研究發現：在 1936 年有財力安裝電話和訂閱雜誌的人不是全體選民的橫截面。這個富裕組群是特殊的組群；這是一個有偏差的樣本，因為大多數樣本是共和黨選民。這樣本選擇 Landon，但全體選民卻不以為然。

基本樣本被稱為**隨機**(random)，在母體群中被選中純粹是偶然；統計人員指全體為「母體群」，樣本是其中部份：索引卡每十個名字選一個，每批紙張取五十張，在鬧市每二十名行人採訪一位。（但請記住，這不是這個國家或城市人口的樣本，只是當時鬧市區域的樣本。一項民意調查的訪問員聲稱可在火車站「找到各種人等。」必須指出她的誤區：例如，帶著小童的母親可能比例不足。）

隨機樣本的測試是這樣的：是否每一個名字或事物在整體中有平等機會成為樣本？

純隨機抽樣[4]，是唯一可以利用統計理論檢查而又令人有全面信心的統計方法，

[4] purely random sample

但其多種用途的成本昂貴和執行困難，令人望而卻步。民意調查和市場研究這些普遍領域幾乎都採用更經濟的替代品：**分層隨機抽樣**[5]。

要得出分層抽樣，先把母總群按已知**盛行率**[6]比例分為**組群**[7]。麻煩從此開始：所知的比例訊息可能不正確。調查員按指示訪問多少名黑人（以收入階層細分百分比），多少名農民等等；這些組群必須均分為四十周歲之上和之下。

聽起來有層有次，但實際情況是怎樣？大部分時間調查員不會弄錯受訪對象是黑人或白人。收入方面會多犯錯。如何界定農民：在農場兼職又在城市上班應如何分類？即使年齡也可能帶來一些問題，避重就輕的辦法是只選擇明顯低於或超過四十周歲的受訪者。在這種情況下，樣本有偏差，沒有包括三十多歲和四十多歲的年齡組。你不能全贏。

考慮以上各點，應如何在分層內得出隨機樣本？最明顯的先找出全體人口的姓名列表，從中隨機選擇；但成本太昂貴。所以訪問員走到街上（偏誤是忽略了留在家中的人們），或是在白天挨家挨戶訪問（偏誤是忽略了上班族），或換到晚上訪問（忽略了電影迷和夜遊人）。

意見調查的操作，歸結到底是對有偏見來源的持久戰，所有著名的民調機構時時刻刻都在作戰。閱讀調查報告時，必須記住這是必然敗北的戰鬥，從來沒有贏過。「英國人有 67%反對…」或其他類似的結果，先要問問這 67%是什麼英國人。

美國著名的人類性學研究者金賽博士[8]與他人合著的《金賽報告[9]》：《男性性行為》（1948 年）及《女性性行為》（1953 年）。《報告》無疑是劃時代的研究，但樣本遠遠不是隨機，令人不安。樣本名單有極大偏差：女性受訪者 75%有大專以上學歷，男性受訪者有頗大比例是囚犯(25%)或男妓(5%)[10]。更嚴重的誤區是樣本大幅度傾向有性暴露狂的受訪者；樂意向訪問員訴說性歷史的人，其經歷大大有異於對訪問員說不的沉默寡言群體。

布魯克林學院 A. H. Maslow 在金賽之前有一項研究，參與的女學生許多後來也志願參與金賽的研究；Maslow 發現這些女生普遍是較為性成熟和獨立特行。這證實了人們對金賽研究的質疑。

閱讀《金賽報告》或任何有關性行為的較近期研究時，要懂得適可而止：即是不

[5] stratified random sampling
[6] prevalence
[7] group
[8] 金賽博士 Alfred Charles Kinsey，1894-1956
[9] Kinsey Reports
[10] 譯文略有補充，參考維基百科。

要過份閱讀。任何基於採樣的研究都突顯這樣的誤區，尤其是大型調查的主要報告濃縮為摘要形式更可能變得如此。

首先，像《金賽報告》這樣的研究至少涉及三個層次的抽樣。上文已指出母體群（第一層次）的樣本並不是隨機，因此可能不特別代表任何母體群。同樣重要的是要記住任何問卷可能只是許多可能問題的其中一個樣本（第二層次）。受訪者的答案只不過是回應那問題的個人態度和經驗的樣本（第三層次）。

類似金賽的性研究和其他調查都發現訪問員的身份會影響調查結果。在二戰期間，美國全國民意研究中心派出兩位員工訪問南方城市的五百名黑人。一位調查員是白人，另一位是黑人。

訪問員提出三個問題。其一是「如果日本征服美國，黑人會得到更好或更壞待遇？」黑人訪問員回報受訪者有 9%回答「更好」。白人訪問員得到同樣的回應只有 2%。黑人訪問員回報受訪者有 25%回答「更壞」。白人訪問員得到同樣的回應卻有 45%。第二條問題以「納粹德國」取代「日本」，結果也是類似。

第三條問題探討可能是基於前兩條問題顯露的感情。「專心擊敗軸心國或致力讓民主更好在美國發展；你認為那一項更重要？」黑人訪問員回報 39%選答「專心擊敗軸心國」，而白人訪問員回報 62%。

偏誤是因為許多未知因素。最有效的因素可能是人們有給出令對方滿意答案的傾向，因此閱讀調查結果時要自我提醒。回答在戰亂時對忠於國家的問題時，南方黑人會告知白人訪問員動聽的答案，而不是本人實際相信的答案，這是不足為奇。也有可能是不同訪問員選擇不同類型的對象接受訪問。

在任何情況下，結果是很明顯是一面倒偏誤，毫無價值。各位可以自行判斷有多少調查的結論是一樣偏頗，毫無價值，而且沒有測試揭示這些偏誤。

如果你懷疑一般調查偏向於特定方向，一如《文學文摘》的錯誤，這可視之為相對證據：受訪者比代表母體群平均組群偏向更有錢，受較多教育，有較多資訊和較高警覺性，更美好的外觀，更常規的行為以及較穩定的習慣。

很容易看到如何產生這此偏誤。假設訪問員被分派到某街角完成面試。眼前兩位仁兄似乎都適合要求的類別：第一位是四十歲的城市黑人，不修篇幅；另一位穿著乾淨工作服，體面整潔。為了盡快完成訪問任務，訪問員更有可能向後者打招呼。全國各地的訪問員都做出類似的決定。

自由派或左翼圈子對民調最反感，普遍認為民調一般被操控。這種觀點的背後事實是民調結果往往不符合那些思想不保守人士的意見和願望。他們指出民意調查似乎選上共和黨，即使此後選民不是這樣投票。

事實上，從上文所見，民調不是必然被操縱，刻意扭曲結果以製造假象。樣本向這一致方向傾斜已是自動扭曲。

補充材料

選擇母體群和抽樣的誤區

書面作業選用那些現有數據？調查選擇那些母體群？全都影響統計數據。

即使母體群的界定符合「涵蓋全體」的意思，如何從中抽樣？[11]

- **簡單隨機抽樣** simple random sampling，也叫純隨機抽樣。從母體群 N 個單位中隨機抽取 n 個單位作為樣本，每一單位有相同機率被抽中為樣本，即是每個樣本單位被抽中的機率相等，每個樣本單位完全獨立，彼此沒有一定的關聯性和排斥性。簡單隨機抽樣是其它各種抽樣形式的基礎，通常只是在母體群單位之間差異程度較小和數目較少時才採用。
- **系統抽樣** systematic sampling，也稱等距抽樣。將母體群的所有單位按一定順序排列，在規定範圍內隨機抽取一個單位作為初始單位，然後按事先規定規則確定其他樣本單位。先從數字 1 到 k 之間隨機抽取一個數字 r 作為初始單位，以後依次取 r+k、r+2k……等單位。這種方法操作簡便，可提高估計的精度。
- **分層抽樣** stratified sampling。將抽樣單位按某種特徵或規則劃分為不同分層，然後從不同分層中獨立、隨機抽取樣本。從而保證樣本的結構與母體群結構比較相近，從而提高估計的精度。
- **整群抽樣** cluster sampling。將母體群的若干個單位合併為組，形成抽樣框，抽樣時直接抽取，然後全部調查中選組群的所有單位。抽樣時只需抽中抽樣框，可簡化工作量，缺點是估計的精度較差。

學術調查較多說明採用那種方法，但一般調查極少說明。以香港為例，有化妝品／牙膏等等廣告標榜「90%（或高比例）女士／牙醫選用…」；為適應法例要求，廣告以極小白字標示數據來自什麼什麼調查。仔細一看，這些調查往往來自內部或母公司調查。這些數據應該是真實的，但這些「內部」調查是否隨機？是否涵蓋適當的母體群？牙醫母體群是否包含全部註冊牙醫，或是參加廣告方主辦免費

[11] 這段落取自〈抽樣〉《維基百科》，略有改寫。

研討會的參加者？「女士」是否局限於在該品牌化妝櫃台瀏覽甚至購物的女士？

- 參考閱讀：抽樣與代表性

輕率概化和過度類化

統計的的特定總體不能代表母體群，即是輕率概化的謬誤，例如調查只限於某政黨黨員和同路人而把結論概化為全民意見。

現實生活中的調查往往以電話進行，常有過度類化的謬誤。如調查人員只致電手機（流動電話），而手機使用者以年青人佔大多數，這忽略了沒有手機，只有家用電話的家庭主婦和老年人。這不是全民調查的正確取樣。

抽樣調查

常見的報導屢屢提到是次調查訪問了多少人。大城市人口動輒千萬，大國人口以億計，究竟調查樣本應有多少才有代表性？不懂統計學的人們少不免懷疑調查數千人是否取得數百萬人的意見。完美公正的抽樣和可信答案的調查，在數學上有誤差範圍，取決於調查的人數。

先要了解取樣調查的兩個重要術語：**置信區間**[12](confidence interval)和**置信水平**[13](confidence level)。置信區間也稱為誤差(margin of error)，即是調查報導時常提到的 ±X%。抽樣誤差本質上不是錯誤(mistake)，最完善的抽樣統計程序和方法都無法避免抽樣誤差（除非剛巧每一個樣本都具有和總體相同的特徵，那另當別論）。

在既定的置信水平，影響其置信區間有三個因素：**樣本大小**(sample size)、**百分比**(percentage)和**母體群規模**(population size)。

很明顯較大的樣本數量更能確保如實反映母體群的答案；也很明顯最大範圍的樣本就是母體群全部，但這是不實際的，否則就無需抽樣調查這回事。但在既定的置信水平，樣本越大，置信區間越少；但這關係不是線性的，不是說倍增樣本大小會導致誤差率減半。

調查的準確度也取決於樣本選取一個特定的答案的百分比。如樣本 99%說「是」，1%說「否」，無論樣本大小，錯誤的機會是微乎其微。然而，如答案的百分比是

---

[12] 亦有譯為「信賴區間」。
[13] 亦有譯為「信賴／信心水平／水準」。

51%對 49%，出錯的可能性要大得多。

樣本可能代表已知的國家或城市人口，或是不確切知道的準車主數目。機率數學證明如樣本是母體群的百分之幾，母體群的規模是無關緊要，除非母體群的規模偏小或是有既定特點的已知群體（例如某協會的成員）。

取樣的黃金規律是「隨機」，真正的「隨機」。調查出錯往往是因為取樣不是隨機。

以大家熟悉的蓋洛普(Gallup)調查為例，看看「美國全國民意調查」是怎麼抽樣的。

> 無論是一次性或追蹤性調查，蓋洛普的取樣是一千人，置信區間為±4%，置信水平為 95%。即使加大樣本，誤差不會有很大差異。
>
> 在收集數據後，蓋洛普依據美國人口調查局的人口特徵（性別、族裔、年齡、學歷和地區）為每位受訪者加權。
>
> 例如，調查一千名國民對總統的支持率為 50%，誤差為±4%，即是支持率是在 46%至 54%之間。如樣本擴大至二千人。誤差可降至±2%，但成本倍增。
>
> 在決定樣本多少時，調查機構必然要考慮成本。最準確的民意調查要涵蓋全體國民，但這是不切實際。
>
> 「置信水平為 95%」的意思是如蓋洛普進行一百次同樣的調查，有九十五次的結果大致相同，只有五次不是在「46%至 54%」的範圍。[14]

---

[14] http://www.gallup.com/poll/101872/how-does-gallup-polling-work.aspx
http://www.gallup.com/poll/File/125927/How%20Are%20Polls%20Conducted%20FINAL.pdf

Sample Size Calculator 是 Creative Research Systems 的網上公共服務，用來決定需要多少樣本以反映目標母體群的精確結果。只要點選置信水平（95%或 99%），輸入置信間距（誤差）和母體群人數，就可以算出所需樣本大小。[15]。

網頁計算器要求輸入以下的選擇，如母體群的規模龐大或未知，可以留空。

決定樣本大小 Determine Sample Size

置信水平 Confidence Level: ( )95% ( )99%

置信間距 Confidence Interval:

母體群 Population:

所需樣本 Sample size needed:

計算置信區間 Find Confidence Interval

置信水平 Confidence Level: ( )95% ( )99%

樣本規模 Sample Size:

母體群 Population:

百分比 Percentage:

置信區間 Confidence Interval:

不恰當的調查問題

問卷和電話調查都是由訪問者提出問題，遣詞用字能引導受訪者給出有傾向性的答案。如二戰期間的民意調查問題為：

- 德國已進佔法國。美國應否參戰？
- 日本已偷襲珍珠港。美國應否參戰？

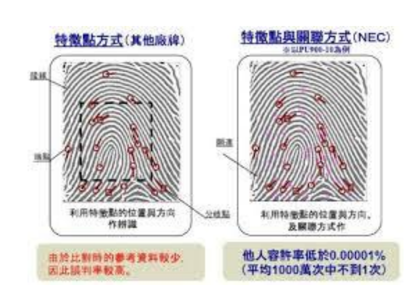
其中的預設立場顯而易見。

另一陷阱是在誘導性提問加入導向「理想答案」的資料。例如：

- 中產家庭稅務是多年新高，你是否支持扣減所得稅？
- 國家提出龐大量赤字預算以應付迫切需求，你是否支持扣減所得稅？

[15] http://www.surveysystem.com/sscalc.htm#one

法律與統計



一宗謀殺官司突顯了嚴重的統計問答。雖然疑犯否認他在犯罪現場，但正面臨控方提出的指紋證據。指紋專家在庭上被控方盤問:「被告人的指紋和其他人的指紋相同的機率是多少？」專家作答:「數十億份之一。」辯方律師盤問:「在犯罪現場得到的指紋被錯誤識別為某人的機率是多少？」專家:「哦，大約是百份之一。」

指紋證據是事實，但識別指紋是判斷，不是事實，是一門科學，並且由機率支配。[16]

〈視頻〉Peter Donnelly: How juries are fooled by statistics 統計如何迷惑陪審團(中文字幕)。統計數字如何錯判「殺嬰案」。

[16] 資料來源：http://www.nytimes.com/2013/05/26/opinion/sunday/when-numbers-mislead.html?_r=0

第二章　精心挑選的平均值

讀者諸君不是勢利小人，我當然不是地產代理。姑且假設你是勢利暴富戶，而我是地產代理。你打算在我熟悉的小區買房子。我打量一下，小心翼翼告訴你這小區的業主住客平均收入每年約一萬英鎊。也許這引起你的興趣；無論如何，你決定買房子，也記住這年收入數目。勢利暴富的你在告訴你的新地址時也不經意拋出這數字。

一年多後，我們又見面了。我是當區地方稅繳納人委員會的成員，要求小區的業主住客簽署請願書呼籲不要增加地方稅或調低物業估值或公交票價減價，理由是這超出小區居民的負擔，畢竟我們的平均收入每年只有£2000。

也許你會附和我和委員會的呼籲；你不僅勢利，也懂得省錢。但你對年收入£2000的說法無法釋懷：究竟我是現在或是去年說謊？

無論怎樣，你不能怪責我。利用統計數據說謊就是這樣的美好。這兩個數字都是合法的**平均值** average，合情合法，都代表同樣的數據，同樣的居民，同樣的收入。都是一樣的。很明顯，至少其中一個是誤導，等同不折不扣的睜眼說瞎話。

我的訣竅是每次拿出不同類型的平均值；「平均值」有非常鬆散的定義。打算影響公眾輿論或出售廣告空間，這一招很管用，有時是無心之失，但往往是故意而為。要清楚明白「平均值」，先要知道是那種平均值：**平均數** mean，**中位數** median 或**眾數** mode。

我拋出一萬英鎊數目時是想提出一個大數值：平均數是這社區所有家庭的收入的算術平均值：所有家戶的收入總和除以家戶數目。中位數是較小的數字：有一半家庭的收入多於£2000，有一半少於這數目。我也可以拋出眾數，這是序列數據最常見到的。如這社區有最多家庭的年收入是£3000，每年£3000 就是眾數。

在這種情況下，沒有解釋的「平均值」是毫無意義；收入數據一般也是這樣。有另外因素亂上添亂：源自隨著某些種類訊息的平均值差別不大，一般來說是無需著意區分。

如果有報告謂某原始部落的男性平均身高只有一米，你會對他們的體型有相當不錯的見解，無需追問這是否平均數，中位數或眾數，三者的數值都是差不多。（當然，如果你打算在非洲出售工作服，就要有比平均值更多的資訊。這是關乎全距

range 和偏差 deviation，下一章詳談。 ）

處理諸如許多人性特點的數據時，不同的平均值是相當接近所謂「**正態分佈**[17]」，以曲線表示其形狀為鐘型；平均數，中位數和眾數都在同一點匯合。

因此，如描述人的高度，各種平均值是一樣好；但如要描述某城市居民的收入，也許是由些微收入至二萬英鎊左右，某地可能有幾個超級大戶。超過 95%的居民的年收入是在五千英鎊之下，曲線向左側傾斜。這不再是對稱的鐘型，而是被扭曲，形狀像小孩的滑梯，梯子急劇上升至一個高峰，滑下部分傾斜逐漸下降。平均數與中間數有相當距離。比對一年的「平均數」和「中位數」，其差異一目了然。

回到上文物業經紀就小區居民年收入拋出兩個相差頗大的平均值，是因為分佈明顯傾斜。如居民大多數是小農戶或打工一族或是年老退休人士，但有三位百萬富翁周末業主，居民總收入的算術平均數是極大數值。幾乎每個居民都在平均數之下。這是現實，但聽起來像笑話或比喻而矣。

因此，讀到企業或東主自白他員工的平均工資是什麼什麼，這數字可能有一些意思，也可能沒有。如數字是中間數，意思是高於或低於中間數工資的員工各佔一半。如果是平均數（如沒有說明，一般是這個），所謂平均收入是£25,000 其實沒有分開東主的得益和和低薪工人的工資。平均年薪£3,800 可能掩蓋工人年薪 £1,400 以及東主以高工資形式拿走大部份利潤。

統計的語言偽術可以把壞事包裝成為較好的外觀。

三位合夥人開設一家小型製造企業。過去一年生意非常好，支付了九十名員工的工資（共£99,000）以及每名合夥人工資各£5,500 後，餘下利潤還有£ 21,000。如何描述這狀況？為便於理解，可以利用平均值。

既然員工都做同樣工作，薪酬沒有太大差別，使用平均數或中位數都是差不多：員工平均工資 £1,100，合夥人平均工資和利潤 £12,500

這看起來很可怕。換一種方式。三位合夥人分取利潤£15,000（餘下£6,000）。這一回以平均數計算員工和合夥人的工資：平均工資 £1,403，合夥人平均利潤 £2,000。

啊！這看起來更好：利潤不足 6%。現在可以發佈，張貼或在談判中使用這些數

[17] normal distribution

據。

這相當粗糙的例子極度簡化，但比對以會計之名做出的花招，這不算什麼。在層次結構和複雜的公司，員工從打字員到年收幾百萬美元獎金的總裁，這樣的手法可以掩蓋各種各樣的東西。

所以，看到平均工資的數字，首先要問：什麼的平均？誰包括在內？美國鋼鐵公司曾表示其員工的平均週薪在不到十年上升了 107%。是的，他們沒說錯—— 但只要注意到十年前的數字包括眾多兼職工人，這數字的意義就大打折扣。如某人去年是半職，今年是全職，他的收入增加一倍，但工資率其實是一樣。

有報導美國家庭的平均收入是$ 6,940。要明白這個數字，先要知道何謂「家庭」以及是什麼平均值。（以及誰這麼說的？他怎麼知道？數字是否準確？）

數字可能來自人口普查局。局方的報告全文說明這是中位數，「家庭」指「住在一起兩個或兩個以上有親屬關係的人」。報告還說明數據來自這樣規模的樣本，每二十個樣本有十九個的估計是在±71 美元的範圍。

這機率和誤差率加起來是相當不錯的估計。調查局人員有足夠的技術和資源以相當精度程度完成取樣研究。想必他們沒有特別要遮掩的。但不是所有的數字都是是在這樣的情況下快樂誕生，也不是伴隨著任何訊息來說明如何精確或不精確。下一章詳解。

看看《時代雜誌》的〈發行人的話〉：新訂戶的年齡中位數為 34 歲，其平均家庭收入為每年$7,270。早前的調查發現舊訂戶的年齡中位數為 41 歲，平均收入為$9,535 美元。問題是為什麼兩次都給出年齡中位數，但刻意沒有說明收入採用那種平均值。

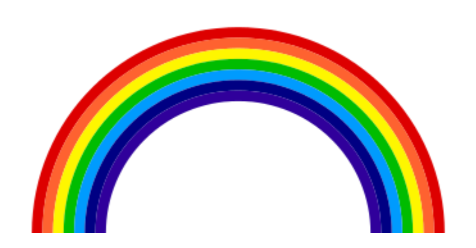
會否是用了平均值以表達較大數值，可以向廣告商介紹讀者群是如此富裕？

利用第一章的耶魯舊生數據，猜猜是採用了那一種平均值。

補充材料

平均值的誤區

討論統計數據時少不免提到「平均值、平均數」。這名詞的表面意思很明顯：平均值就是大致居中的一個數值。但實際上有好幾種平均值。



平均而言，彩虹是白色的。

※**算術平均值**(mathematical average/mean)是把所有數據加在一起，再除以總體的樣本量計算。（3,3,5,4,7）這幾個數值的算術平均值就是把總和(22)除以 5（因為有 5 個數值）；算術平均值是 4.4。

※**中位數**(median)是一組數值從低到高排列，恰好處在中間位置的那個數值。同上例子（3,3,5,4,7），中位數是 4，因為有兩個數值(3,3)比它小，兩個數值(5,7)比它大。

※**眾數**(mode)是一組數值中最常見的數值。同上例子的眾數是 3，因為出現了兩次。

算術平均值看起來似是以上三種計算方式最簡單的一種，但實際上不是這樣。因為一組數據中如有過高或過低數值（極端的數值）對算術平均值產生很大的影響。

※例如，統計一個社區內 50 戶家庭的收入。大多數家庭的收入是每年 $40,000-60,000，但有一家每年收入是 5 百萬元。如此這般的算術平均值因為 5 百萬元這個數值而大大提高。

※如 9 個人各有 1000 元存款，第十個人只有 1 元存款，算術平均值是 900.10 美元。

比較可信的數據調查往往去掉最高和最低的數值才計算算術平均值。但不是每一

項調查都這麼可信。除非看到所有數據或已去掉極值的說明，最好不要對這些數據照單全收。

中位數的誤區



中位數容易有誤區，因為和其他數據相比，這不是很明顯過高或過低。中位數處於中間位置，很容易隱藏了那些很大或很小的數值。例如，數據是 0.1,1,2,3,4,5,3000，中位數是 3。

用中位數描述某事件隨時間變化的程度時，容易遮掩事實。如過去九年每年漲價 3%，但今年漲價 20%，中位數仍然是 3%。

如總體樣本數量是偶數，計算中間兩個數值的平均值作為中位數，可以避免極值的影響。

眾數的誤區



如數據組龐大，較少機會出錯；如數據組較小，容易有誤區。

※例如，如數據組數值都在 1-100 之間，但 1 出現了 3 次，那麼 1 就成為眾數，雖然平均值（這種情況下比較敏感）會接近 50。

※大規模調查可以通過強調眾數來操控。100 受訪者對某產品的滿意度在 1-10 之間打分，即使打 10 分的人數比其他分的人數只多了 1 個，10 就是眾數。

- 〔視頻〕算術平均數、中位數、眾數之比較（國語）
- 〔參考〕算數平均數，中位數、眾數

想一想〈五個整數〉

有五個整數，其平均數是 4，眾數是 1，中位數是 5。求該五個整數。

解題及答案

既然眾數是 1，必然最少有兩個整數是 1。因為中位數是 5，第三個整數必然是 5。這個數字組是{1, 1, 5, x, y}。

如平均數是 4，五個整數的總和必然是 4 x 5 =20；即是 1 + 1 + 5 + x + y = 20，暗喻 x + y = 13。

以下說明最簡單的情況：假設 x 是少於或等於 y，如 x = y，得出 x + x = 13, 2x = 13, x = 6.5。明顯 x 是大於或等於 5，因此 5 是少於或等於 x 少於或等於 6.5。

因此，如 x = 5 就會有兩個眾數：1 和 5。因此可推論 x = 6, y = 7，而這五個整數必然是{1, 1, 5, 6, 7}。

資料來源：http://mathschallenge.net/full/average_problem

第三章　不存在的小數字

一位統計學家建議，看到一項調查結果時就要質疑：「前後有多少個陪審團才找到這一個？」

如前所述，採用頗為偏差的樣本可以產出幾乎任何結果；依常規的隨機採樣，如規模小而又多番使用，也可以產生幾乎任何結果。

「用家改用白齒牌牙膏後，蛀牙減少 23%！」仔細閱讀，說明還聲稱調查結果來自令人放心的「獨立」實驗室，數據也是由特許會計師認證。還要什麼更多證據？

然而，大多數人從經驗中知道什麼牌子的牙膏都是差不多。為何白齒牌的用家有這樣的聲明？這廣告是否說謊？沒有，況且廣告不必說謊。有更簡單更有效的方法。

第一個攪局的因素是樣本不足，不符合統計學的要求。廣告的小字說明測試組群只有十幾人。[18]

有些廣告會忽略這訊息，即使精通統計也只能猜想這是什麼品種的詭辯。在類似的情況，十幾人的的樣本不是那麼糟糕。幾年前，有一種牙粉上市，自稱「矯正齲齒相當成功。」當時的想法是該牙粉含有尿素，已由實驗室證明有效。這是毫無意義的，因為這初步試驗只涉及六個案例。

那麼白齒牌牙膏沒有說謊，又如何得出被認證的結果？讓任何小組樣本在半年內記錄蛀牙數目，然後改用白齒牌牙膏。只有三個必然的結果：蛀牙明顯更多、明顯更少或沒有明顯變化。如果是第一或第三個情況，白齒牌牙膏把數據存檔（在看不見的地方）並重覆調查。遲早，只是因為機率的操作，測試組必然出現第二種情況，值得大吹大擂，作為廣告標題。無論測試組是用蘇打或其他牙膏，都會出現第二種情況。

利用小組群的重要性是這樣的：在大組群機率產生的任何差異很可能只是少許，不值得大書特書。減少蛀牙 2%的廣告不會讓牙膏大買特買。

小規模樣本只憑機率產生的變化，實在不能說明什麼。來一個小實驗吧。

---

[18] 譯註：許多國家的保護消費者法例要求廣告說明調查的主辦方，日期和樣本數目。

人人都知道拋硬幣花紋朝上的機率是一半一半。拋硬幣十次，花紋朝上的可能有八次，這「證明」花紋朝上的機率是 80%。牙膏統計就是這樣。只拋幾十次，有可能得出 50%的結果，但不大可能。但是，如果耐心拋上一千次，幾乎可能極為接近 50%（但不完全肯定）的結果；這才是真正的機率。要有相當數量的測試，平均規律才可以是有用的描述或預測。

多少次測試才算足夠？這是棘手問題，取決於受採樣調查的母體群其數量和其中差異的程度。有時，樣本的數目並不是表裡如一。

幾年前有一個顯著的例子是關於脊髓灰質炎疫苗的試驗。這似乎是一個令人印象深刻的大規模醫學試驗：450 名兒童接種疫苗，對照組是 680 沒有接種的兒童。此後不久，社區爆發流行病。曾接種疫苗的兒童沒有一人感染小兒麻痺症。

但對照組的兒童也沒有感染。在設計試驗時，相關人員忽視或不理解麻痺性脊髓灰質炎的發病率較低。以一般發病率計算，這規模的母體群只預期有兩宗病例。因此這測試從一開始就注定沒有意義。測試母體群要有十五或二十五倍的規模才可以得出稍有意義的答案。

許多偉大的醫學發現曾在類似的情況下急急出台。正如一位名醫所說：「要趕快採用新醫療措施，以免為時過晚。」[19]

犯錯的不限於醫學界。公眾壓力和草率報導往往迫使未經證實有效的治療提前發動，尤其面對當前龐大需求而統計數據朦朧不清。幾年前的感冒疫苗和近年的抗組織胺藥就是例子。這些失敗的「靈藥」之深受歡迎，主要是因為疾病的不可靠本質和邏輯的缺陷。感冒無需吃藥，過幾天就會自我治愈。

如何避免被不確定的結果愚弄？不可能人人是統計學家懂得研究原始數據。有一個很容易理解的顯著性檢驗：究竟報告的測試數字有多大可能是真實的結果，而不是偶然產生。這是非專業人士不明白而且不存在的小數字。

如訊息來源有給出**顯著水準**[20]，就更容易掌握。顯著水準最簡單的表達方式是機率。人口普查局給出「機率為 19/20」，表明具體的精確度。在大多數情況下，這 5%顯著性水準已經夠好。有一些較嚴格的要求 99/100 的機率，這意味著確切顯著差異機率為 1%，這有時被描述為「實際肯定」[21]。

[19]傳聞這句話出自 William Osier 爵士和 Edward Livingston。他們都同是醫生和這方面的權威。
[20] degree of significance
[21] practically certain

還有另外一種可能同樣有害的不存在小數字。這小數字說明事件的範圍或其與平均值的偏差。平均值（無論是平均數或中位數，具體或不具體）往往流於過於簡化，比無用更糟糕。一無所知通常好於一知半解；只知皮毛可能是危險的事情。

例如因為統計數據家庭有三至六人，據此規劃建房，房子有兩間臥室供三至四人居住。這「平均」規模的家庭實際上只是家庭總數的少數。為「平均」家庭建造房子，而忽視人數較多或較少的家庭；一些地區已經有過多兩間臥室的房子，而較小和較大的單位不足。這誤導而又不完善的統計已導致代價高昂的後果。公共健康小組指出：「算術平均值歪曲了實際的情況：三人和四人家庭只有 45%。35%是一人及二人家庭，20%是四人以上。」

人們面對「三至六人」的權威數字，莫名其妙地失去理智，抵消了人們從觀察中得知的印象：很多小家庭，少許大家庭。

類似的不存在小數字情況是令無數父母擔心的所謂「格塞我常模[22]」。家長在週刊和報章讀到小孩三個月大學會坐起來，立即就想到自己的小孩。如小孩三個月大還沒有坐起來，家長往往得出結論小孩是「弱智」或「不正常」等等令人反感的顧慮。由於小孩必然有一半到了三個月大不會坐起來，很多父母不開心。當然，從數學上來說，有另一半的父母發現自己的小孩「勝於他人」，他們的喜悅平衡了前一半父母的憂愁。如憂愁的父母強迫小孩符合常模，會適得其反。

這一切並不是說 Arnold Gesell 醫生和他的方法有什麼問題。問題出自聳人聽聞或學藝不精的作家過濾了研究人員的訊息，未有留意在這過程中消失了的數字。如果這些「常模」或平均值能補上正常範圍的說明就可以避免很多誤解。父母看到自己的小孩是屬於正常範圍，不會擔心那些微小而無意義的差異。幾乎沒有人在任何方面是完全正常，就像拋硬幣一百次很難會得出五十次是花紋向上。

混淆了「正常」與「理想」讓這一切變得更糟糕。Gesell 醫生只是簡單說明一些觀察到的事實；只是擔心的父母在閱讀書籍和文章時以為小孩坐起來比常模慢了一天或一個月必然是比別人遜色。

對金賽性學博士的大多數愚蠢批評（其實很少人曾透徹閱讀）來自把「正常」等同良好，優異，可取。金賽博士被指控把各種常見但不受認可的性行視為正常，因而荼毒青年人心靈，向他們灌輸有害的思想。但他只是陳述他認為這些是正常活動；這正正是「正常」的意思，他沒有加上任何「認可」的印章。他不認為他是判斷這些行為是否「不可取」的權威。博士碰上了一直困擾著許多其他觀察員的危險難題：提出任何情感敏感的內容而不另行草草陳述你是否支持或反對。

[22] Gesell's norms

不存在的小數字其欺騙性不是因為沒人留意這不存在，雖然這是小數字成功的秘訣。現今對新聞工作者的批評是譴責「坐在辦公室的記者」不再如老派記者去「跑新聞」，而是不加批判地重新編寫政府的新聞稿。以下的不思進取新聞樣本來自新聞雜誌《雙週刊》〈工業新發展：西屋公司冷浴法增強鋼硬度三倍〉。

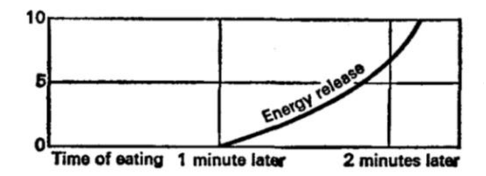
這聽起來像不錯的發展，直到讀者試圖明白這是什麼意思，這句子變得難以捉摸。新浴法是否在處理後增強鋼硬度三倍？抑或生產的鋼鐵其硬度是三倍以前的任何鋼鐵？冷浴法有什麼作用？看來，記者只是傳遞文字，沒有探討其中意思，而是期望讀者水過鴨背，看過了就以為快樂地學懂一些什麼。這讓人聯想到課堂教學講授法的舊定義：教師把教科書內容傳送到學生的筆記本電腦，雙方都沒有動腦筋的一個過程。

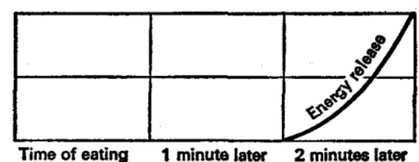
幾分鐘前，我尋找《時代》周刊一些關於金賽博士資料時，發現另一不堪細看的語句。這是電力公司在 1948 年的廣告：「時至今日，超過四分之三的美國農場有電力可用」。這聽起來很不錯。這些電力公司真的很賣力。當然，小心眼的可以意譯為「幾乎四分之一的美國農場沒有電力可用」。但是，真正的噱頭是「可用」這個詞語；電力公司利用這詞語自說自話。明顯地這並不意味著所有這些農民實際上用上電力；若然是這樣，廣告肯定會明確說明。所謂「可用」可能只是意味著電線掛在農場的上空或是十或百里的距離。

這是流行雜誌一篇文章的標題：〈現在可以預測你的子女將來有多高〉。文章的顯眼處展示一對圖表：一個是男孩，一個是女孩，顯示孩子成長期的身高會是最終身高的比例。「要確定孩子成長後的身高，核對現在的測量高度。」

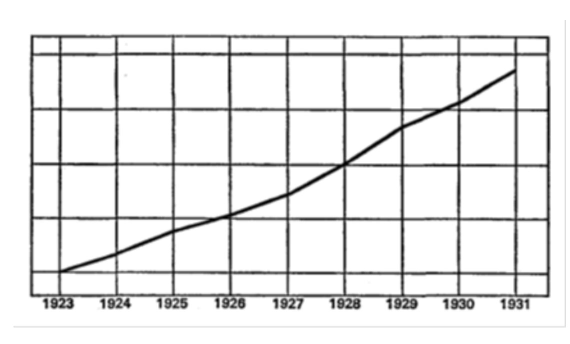
這文章和圖表的致命弱點是忽略了不是所有孩子都是以同樣的方式長高。有些慢慢長高後加快，有些突然長高一段時間然後趨於平穩緩慢，還有一些是相對穩定的長高。這些是基於大量測量結果的平均值。以總數或平均數計算，隨機取樣一百名年輕人的高度這毫無疑問是準確的，但父母感興趣的只在某時刻的高度，這樣的圖表幾乎是一文不值。想知道孩子將來會有多高，觀察他的父母和祖父母可能得出更好的猜測。這不是很科學和準確，但至少比圖表準確。

我十四歲時參加高中軍訓班，按身高排在矮子班，按圖表我最終身高應該是 5 英呎 8 英吋。現在我是 5 英呎 11 英吋。預測身高有三英吋的錯誤是極為差勁的。





有兩盒葡萄+堅果+麥片的早餐食品，不同的包裝，都有「科學家證明這是真的！」的圖表標榜「在兩分鐘內開始給你能量！」左圖表在左邊列出數字，右圖省略了數字。數字沒有說明代表什麼，沒有意思；反正兩個圖表都沒有特別意思。圖表顯示陡峭的攀爬線，分別顯示在進食後一分鐘（左圖）和兩分鐘（右圖）後能量釋放。左圖的能量線爬升約快一倍，這表明繪圖人員沒有想到這些圖表是什麼意思。



這種愚蠢圖表可能只是想吸引青少年或早上半夢半醒的疲憊家長。沒有人會用這樣的統計圖來侮辱大商巨賈的智慧吧…或者會吧？《財富》雜誌的廣告宣傳欄經常刊載某機構業務逐年上升趨勢的令人印象深刻圖表。圖表沒有數字。究竟這是業務增加一倍或一年逐年以數百萬美元增加，或是以蝸牛速度每年只增加一兩元，不得而知。

如平均值或圖形或趨勢沒有包含一些重要數字，就要加倍小心。露營人士不會依賴平均溫度的報告選擇營地。61℃是舒適的平均溫度，在加州的可選範圍包括內陸沙漠和海岸離島。但中間數忽略了範圍：內陸沙漠的溫度範圍 15～104℃，海岸離島是 47～87℃。

第四章　為了子虛烏有無事忙

Josiah Stamp 爵士記述 Randolph 勳爵研究收入的報告。他的私人秘書一直站在旁邊。勳爵說：海關收入比去年同期增長 34%，令人欣慰。秘書糾正他，指出這只是 • 34%。

「這有什麼區別？」勳爵問道。祕書解釋 34 是 • 34 的一百倍，勳爵說：「我經常看見那些該死的小點，但從來不知道他們的意思。」

小數點和其他該死的差異突然出現，困擾著測試成績的比較。不介意的話，提一個例子。國光和美蓮參加智力測驗。很多學生在求學時期都會參加類似測驗，已成為這個時代的主要巫術偶像之一，可能要爭論要花功夫才能找出測試的結果；訊息是如此深奧，經常被認為要交由心理學家和教育學家處理才是安全的。無論怎樣，國光測試的智商是 98，美蓮是 101。當然，智商是基於 100 的平均或「正常」水平計算。

啊！美蓮是聰明的，高於平均水平；國光低於平均水平。不要糾纏於這些結論，因為任何這樣的結論都是無稽之談。

先要說清楚：無論智力測驗計量的什麼東西，並不是我們一般以為的智力。智力測驗忽略了一些重要的事情，例如領導力和創造性的想像力，沒有考慮到社交場合的判斷能力，或是音樂、藝術或其他能力的傾向，更不要說努力處事和情緒平衡等性格特徵。最重要的是學校最經常給出的測試是閱讀測試（快速和便宜）；慢讀的學生不可能拿高分。

假設我們已經認識這一切缺點，並同意智商僅僅只是計量一些定義含糊，處理抽象問題的能力。也假設國光和美蓮參加的是一般認為是最好的個別測試，並且不要求任何特定的閱讀能力。

智商測試聲言是智力的採樣。一如任何其他抽樣方法的產品，智商是一個有統計誤差的數字，誤差影響智商數字的精確度和可靠性。

這些試題就像隨機在農田採摘玉米，採摘了一百條玉米，應當對這塊農田的種植狀態心中有數。這樣的訊息已足以和其他玉米田比較（如兩塊玉米田不是很相似）。如兩塊農田差別不大，可能要採摘更多玉米，並以一些確切的質量標準評價採摘的樣本。

玉米樣本能如何準確代表整塊農田，可以用可能誤差和標準誤差[23]的數字表達。假設要在柵欄以外目測許多農田的大小，第一件事可能是先測量步行一百碼的誤差。如經多次步測，發現誤差的平均值是三碼，即是說步測有一半是超出三碼，一半是少了三碼。

那麼能誤差是每一百碼有三碼，或 3%，因此記錄步測結果是 100±3 碼。（大多數統計學家現在更喜歡用另一種但相等的標準誤差[24]，只算計約三分之二的事件，而不是一半半，在數學計算方面更為方便。本書集中在可能誤差，Stanford-Binet 測驗也是這樣使用。）

一如以上的步測例子，Stanford-Binet 智商測驗的可能錯誤已證實為 3%。這不是關乎測驗的優劣，基本上只是表達測驗是否一致。所以國光的智商可以更充分地表達為 98 ± 3，美蓮是 101 ± 3。

這是說國光的智商是在 95～101 的範圍，他在這範圍內可能是高於或低於任一智商數字，機會均等。從而可見美蓮的智商高於或低於 98～104 範圍任一智商數字的機會也是均等。國光智商高於 101 有 1/4 機會，美蓮的智商低於 98 也是有 1/4 機會。有 3%以上機會國光不是遜色，而是優異。

這歸納為解讀智商和許多其他採樣結果的唯一方法是在範圍之內。「正常」不是 100，而是 90～80（舉例而言），也就是說比較在這範圍內和在較低或較高範圍的兒童才有一些意義。比較只有極小差異的數字是沒有意義。必須始終記住這 ± 符號，即使（或尤其是）沒有特別說明。

無視這些隱含在所有採樣研究的誤差，只會導致了一些極為愚蠢的行為。有雜誌編輯奉讀者調查為福音，主要是因為他們不理解。男讀者有 40%偏愛一篇報導，只有 35%喜歡另一篇，他們要求更多類似第一篇的報導。

對雜誌來說，讀者的 35%和 40%之間的差異可能是重要的，但調查中的差別可能不是真實的。為了節省成本，讀者樣本往往減少到只有幾百人，尤其是淘汰了那些誰根本不看雜誌的人們。主要吸引婦女的雜誌其男讀者樣本的數目可以是非常小。這些再細分為「閱讀全部文章」，「閱讀大多數文章」，「閱讀一些文章」和「不看文章」各分類，那 35%的結論可能只是根據極少樣本。隱藏在這些數字背後的可能誤差會是如此之大，依賴這結論的編輯等同瞎子摸象。

[23] probable error and the standard error
[24] standard error

有時，人們為了一些數學上是真實和顯著但是如此微小以至沒有意義的差異而大費周折。這違背了古語的智慧：「差異如會導致差異才是差異」。一個典型例子是「老金牌」香煙為了一些子虛烏有的事情而吵吵鬧鬧，並從中獲利。

《讀者文摘》的抽煙編輯無意中開始這場鬧劇。他們本來認為所有牌子的香煙都是一樣的。雜誌委托實驗室分析幾個牌子香煙的濃煙，並公佈結果：全部牌子香煙的尼古丁和諸如此類東西的內容。雜誌詳列詳盡數字，證明所有牌子的香煙實際上是相同的，抽那一個牌子沒有任何區別。

你可能認為這是對捲菸製造商和構思新廣告角度的廣告公司是一大打擊，這似乎完全推翻了香煙舒緩喉嚨和對人體無害的廣告聲言。

但有人發現在幾乎相同毒素含量的列表中，有一牌子的香煙必然排名最低；這就是「老金牌」。於是報章出現了最大標題的廣告，標示這本全國通行的雜誌測試所有香煙，「老金牌」含有最少數量的不良物體，但剔除了這些差異可以忽略不計的說明。最後，「老金牌」被責令終止這種誤導性廣告。這並沒有任何影響；「老金牌」已從中得到好處。

補充材料

以會員制組織的公司討論業積。營銷部門的統計顯示上月的新會員人數是全年最高。這只是部分正確。翻查記錄，前兩個月的退會人數也是整年最高，會員人數基本持平。上月的新會員人數也是與去年同期相若，表明這不是新趨勢。[25]

[25] 資料來源：http://zestsms.com/about/blog/statistically-irrelevant/

第五章　嘖嘖稱奇的圖形

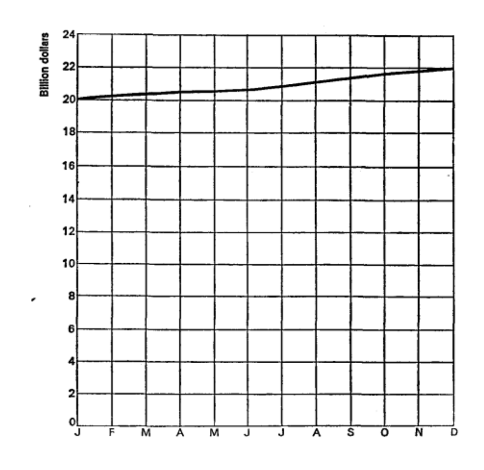
數字是恐怖的。小矮胖信心滿滿告訴愛麗絲，他是文字的主人；但許多人對數字沒有同樣的信心。也許這要回溯我們早期數學經驗導致的創傷。

不管是什麼原因，這對於渴望讀者眾多的作家，計劃廣告能多賣貨物的公司，期望書籍或雜誌大受歡迎的出版商，這確實是一個真正的問題。常見的情況是表格形式的數字是禁忌，文字又未能充份表達，往往只有一個答案：插圖。

最簡單的統計插圖，或**圖形** graph，是不同的線條，用於顯示趨勢很有用，實際上大家都有興趣利用圖形去知道或表達或指出或譴責或預測。

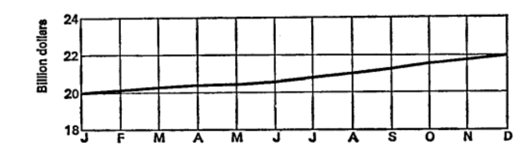
以下圖形顯示國民收入如何在一年之內增加 10%。

先劃出方格，底線寫下月份，左邊標示「以十億元計」。在方格標出數據點，連起來完成圖形：



這很清楚，表明年內發生了什麼，並且標明每個月的升幅。人人容易理解，因為整個圖形是按比例，而且底線有 0 值作為比較。10%看來就是 10%：上升趨勢是實質的但也許不是壓倒性。

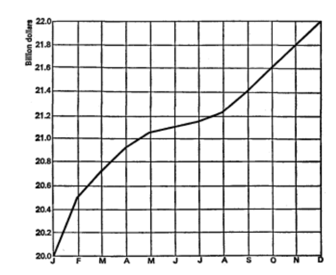
如果只是想傳達訊息，這是非常好。但是，假如想贏得爭論，震撼讀者，促使他轉化為行動，賣東西給他，這圖形不夠誇張。斫掉底部。



這更像樣了。（也減少用紙；這是向挑剔人士反對這誤導性圖形的好理由。）數字相同，曲線也相同，圖形也相同。沒有什麼是偽造的 - 除了給出的印象。匆促的讀者只看到國民收入線十二個月爬升了一半的篇幅，這是因為已經不見了被裁掉的部份圖形。一如語法課中的缺失句子部分，這是「不言而喻」。當然，眼睛不「理解」不存在的東西；小小的增長在視覺上成為大大的增長。

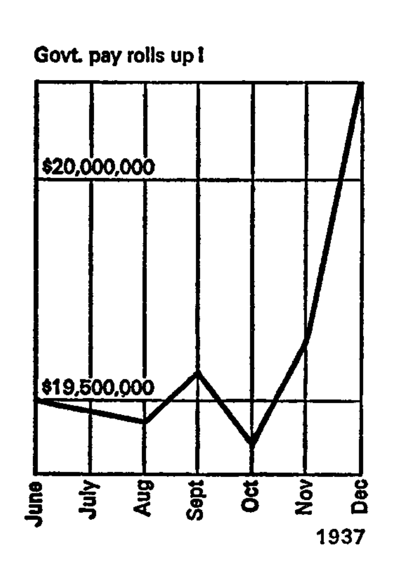
既然練習了欺騙，為什麼停下來？還有進一步的伎倆可用，讓微薄的 10%看起來

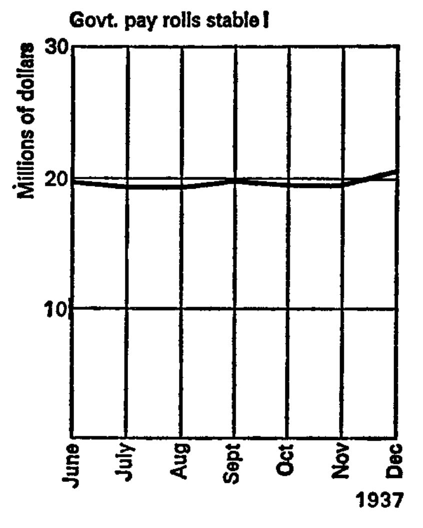
更活潑有力。簡單地改變縱坐標和橫坐標之間的比例。沒有任何規則反對這樣做，並且給出更漂亮的圖形。要做的只是把縱坐標答比例從 2 元改寫為 0.2 元。



這令人印象深刻，是不是？讀者會感到全國經濟繁榮。這是改寫「國民收入上升 10%」為「國民收入急增 10%」。這更有效，因為沒有包含任何形容詞或副詞破壞客觀性的幻想。沒有人可指責你。

這樣的例子不止一個。一份新聞雜誌用同樣方法顯示股市創下新高，圖形被截斷，以使看起來攀升得更利害。哥倫比亞天然氣公司的「我們新年度報告」的重刊圖表。如果仔細閱讀和分析小數字，會發現十年內生活成本上升約 60%，而天然氣的成本下降了 4%。很不錯，但顯然哥倫比亞天然氣認為還不夠好，於是在 90%砍掉了圖表（沒有縫隙或其他警告指示）。所以，讀者見到的是：生活成本增加了兩倍多，天然氣成本下降三分之一！





政府薪資大幅增加！　　政府薪資平穩

鋼鐵企業曾使用類似的誤導圖形試圖影響輿論反對工資上漲。這不是新手法，很久以前已有這樣的不當行為，不僅只是在統計學專業期刊。《鄧氏評論》主筆早在 1938 年看出左圖的破綻：標題是「政府薪資大幅增加！」，曲線從底部急升至頂部，使得增加 4%的樣子看來超過 400%。右國是修正圖形：給出了相同的數字，誠實的紅線僅上漲了 4%，標題改寫為「政府薪資平穩」。

補充材料

圖形的誤區

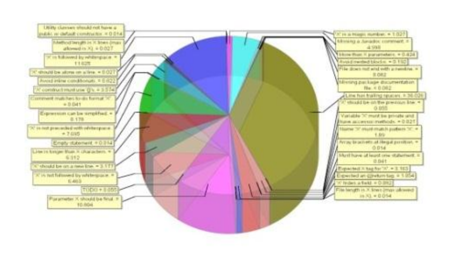
在統計學中，誤導圖形也稱為扭曲圖形，歪曲了數據，構成統計誤用，導致不正確結論。

圖形誤導可能是因為過分複雜或製作粗糙，但精心泡製的圖形也可以導致不同解釋。誤導性圖形可能是故意，以隱瞞數據；或是無心之失：錯用了繪圖軟件，錯解數據，或是數據不適合圖形表達。〔虛假〕廣告特多用上誤導性圖形。

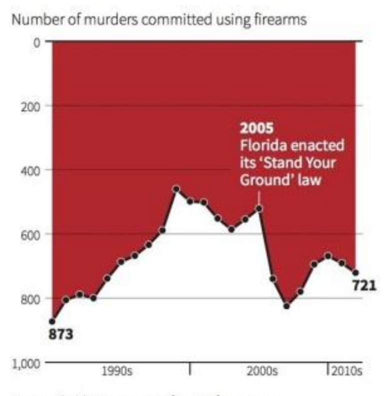
美國統計學家 Edward Tufte 創造了「垃圾圖表 chartjunk」這個新字：

> 「圖形的室內裝修佔據大量篇幅，但沒有告知讀音什麼新的東西。裝飾的目的各不相同 - 使圖形看起來更加科學和嚴謹，使表達顯得活潑，讓設計師有機會展現技能。不管其原因，這些篇幅都不是數據或只是冗餘數據，並且往往是 chartjunk。...Chartjunk 可以把沉悶數據變得慘不忍睹，但不能遮掩數據之不足。」[26]

不當使用圖形



不需用圖形而使用圖形可能導致不必要的混亂／解釋。一般情況下，圖形要配上越多解釋，這圖形的實際需求其實越少。圖形表達不總是比列表更好表達訊息。[27]

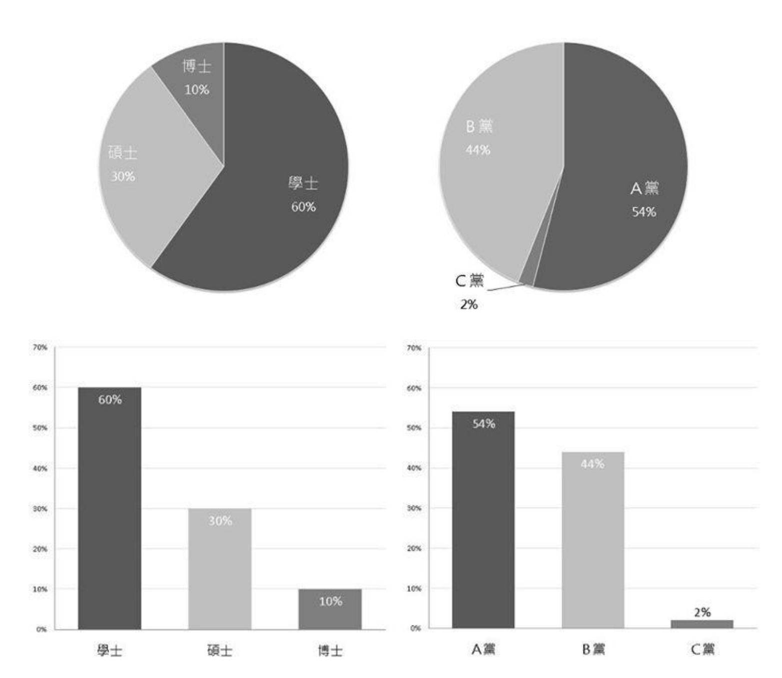


偏頗的圖形

偏頗的圖形標題，標籤或標題不恰當地誤導讀者。左圖是美國佛羅里達州因槍擊致死的統計圖形。驟眼看來，在 2005 年訂立「市民自衛法」後，槍擊致死事件從高位回落。仔細一看，這圖形違反一般常規，直軸是從 800 倒數至 0！數據是真實的，但嚴重誤導。[28]

[26] *The Visual Display of Quantitative Information.*
[27] 插圖取自 http://www.theusrus.de/Blog-files/pie_chart.jpg
[28] http://www.livescience.com/45083-misleading-gun-death-chart.html

圓形圖的誤區



圓形圖最重要的功能在於呈現整體中各部份的組成和比例。其實條形圖(bar chart)更適合比較各個組成部份的差異；雖然讀者熟悉時鐘角度，但還是比不上對於長度的感受。如果不看數字，條形圖比較容易看出學士人數是碩士的兩倍，碩士是博士的三倍。[29]

Edward Tufte 在有這樣的說法：

> 「表達小的數據集，列表比圖形圖好很多。列表幾乎總是優於愚蠢的圓形圖；唯一比圓形圖更糟糕的是幾個圓形圖，因為讀者要在多個圖形之間的混亂空間要作出比較。圖形圖的數據密度低，又不能在視覺層面把數值排序，因此不應該使用。」[30]

---

[29] 這一段和下一段以及黑白插圖取自〈圓形圖的使用〉，略有改寫。

[30] *The Visual Display of Quantitative Information* p.178

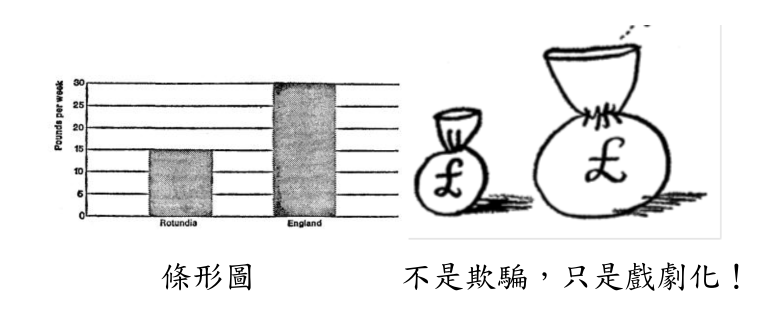
第六章　一維圖形

上一代時常提到「小人物」，即是所有的人。這聽起來太白鴿眼，我們成為「老百姓」。這也很快被遺忘，現在我們是「國民、公民、市民」。但「小人物」依然存在；他就是圖形上的人像。

圖形選擇形象化，以一個小人代表一百萬人，一個錢袋或一堆硬幣代表一千英鎊或一百萬美元，一塊牛排代表明年的牛肉供應；這些全是**圖形統計圖表**[31]，一種有用的設備，吸引注意，也能夠成為流暢，狡猾和成功的騙子。

圖形統計表源自普通條形圖[32]，用於表達和比較兩個或兩個以上數據的簡單和流行方法。

條形圖也能夠瞞騙。如圖形只表達一個因素，但改變了條形的寬度和長度，或以體積難以比較的三維物件代替條形，這圖形值得懷疑。被截斷的條形圖一如被截斷的線形圖同樣的啟人疑竇。地理書，公司聲明和新聞雜誌往往用上條形圖，也用上吸引眼睛的圖形統計圖。



條形圖　　　　不是欺騙，只是戲劇化！

如目的在於溝通訊息，條形圖已可滿足要求。但我想要更多。我想說的是英國工人的待遇遠遠比 Rotundian 更好，我越能戲劇化表達￡15 和￡30 的區別，我的論點越引人注目。說實話（當然我不打算這樣做），我希望你從圖形推斷出一些東西，讓你得到誇張的印象，但我不想被你看破我的招數。有一種方法，而且每天都有人這樣欺騙你。

我只是畫一個錢袋表示 Rotundian 的￡15，又畫一個大一倍的錢袋代表英國人的￡30。這是按比例的，是不是？我追求的是你的感覺。英國工人的工資遠遠多於外國人。

[31] pictorial graph or pictograph
[32] bar chart

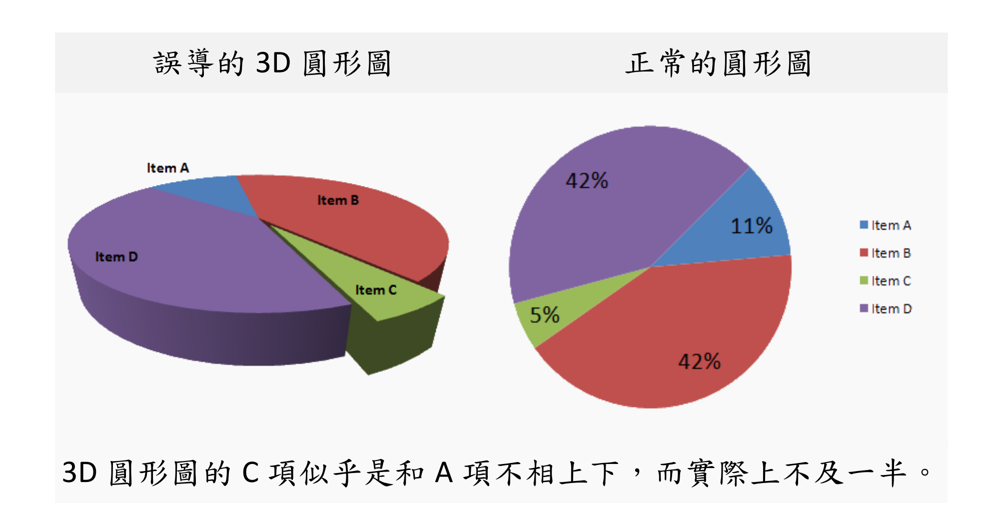
當中的詭計是這樣的。因為第二個錢袋是第一個的兩倍高和兩倍寬，佔用篇幅不是兩倍，而是四倍。數字依然是二對一，但佔據主導地位的視覺印象是四比一，或者更多。因為這些三維圖像是立體的，第二個錢袋的厚度必然是第一個的兩倍。幾何教科書指出類似立體的體積隨著任何維度的立方而改變：2x2x2=8。如第一個錢袋有￡15，第二個應有￡120。

那確實是這巧妙小圖給出的印象。雖然是說「兩倍」，我實際留下了八比一壓倒性比例的持久印象。

你也很難指責我我有任何犯罪意圖。我只是隨波逐流。新聞雜誌反復這樣做，一如上例的錢袋。

補充材料

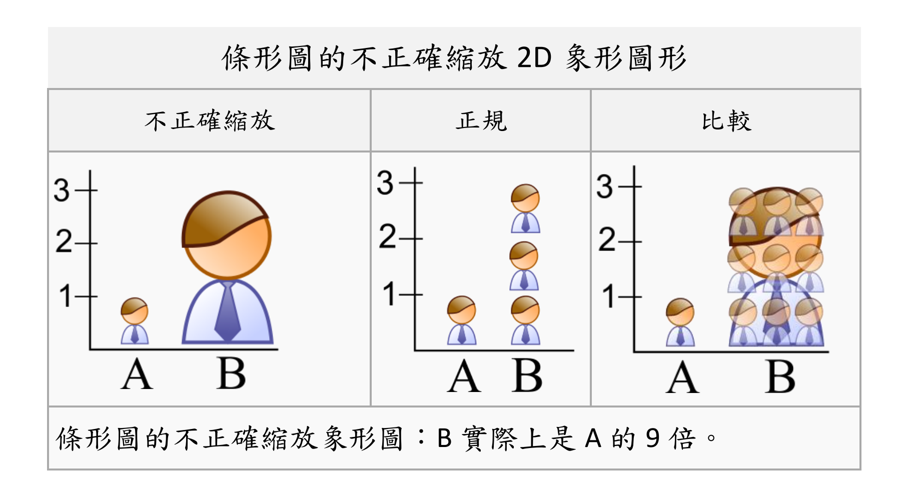
很多統計圖形不適合三維(3D)形式，圓形圖特別如此。由於消失點效果，即使同樣大小，3D 圓形圖靠近讀者的部份會看起來比較大塊，較遠的部份比較小。這扭曲了資料的呈現。只是為了美觀而犧牲精準表達，說不過去。下面的例子說明這現象：

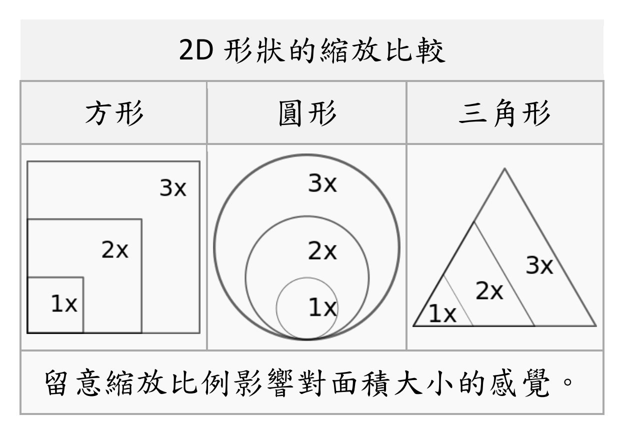


3D 圓形圖的 C 項似乎是和 A 項不相上下，而實際上不及一半。

不正確的縮放

條形圖使用象形比例，不應均勻縮放，因為這導致誤導性比較。讀者看到的是象形圖的面積，而不是高度或寬度，導致比例以平方面積解讀。



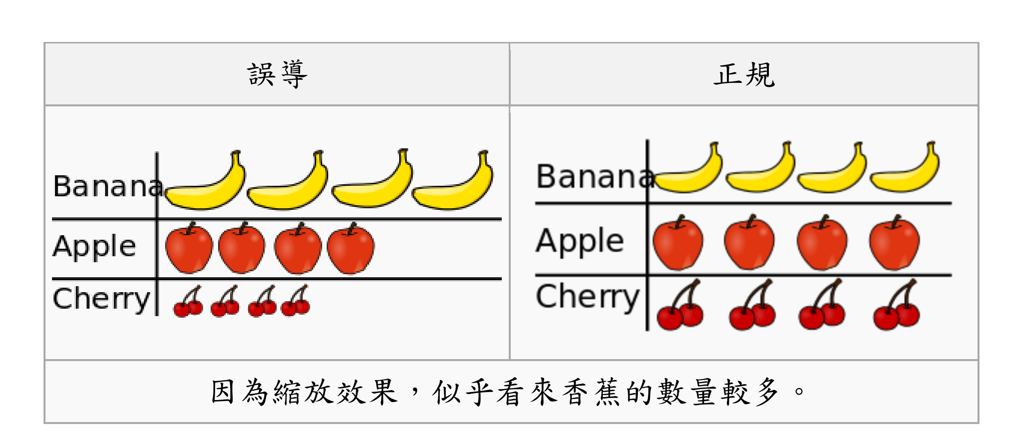


3D 象形圖不當縮放導致立方效果。

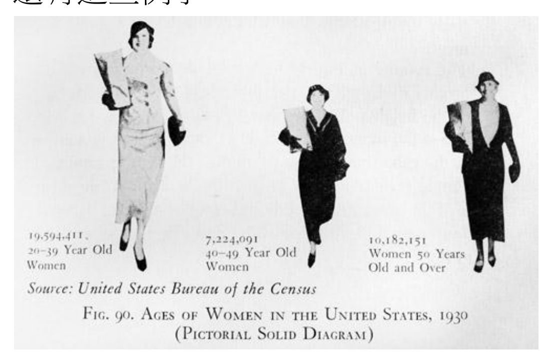


這 3D 象形圖顯示 2001 年房屋銷售比去年有增長。因為沒有直軸說明，讀者無法理解變化；兩倍的縮放看來是八倍(2^3)。

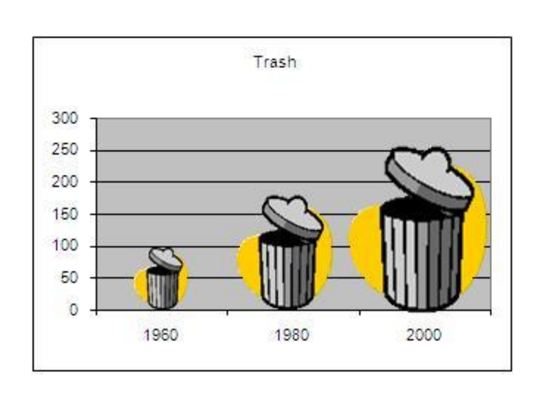
不當縮放的 3D 象形圖誤導讀者以為項目實際上改變了大小。



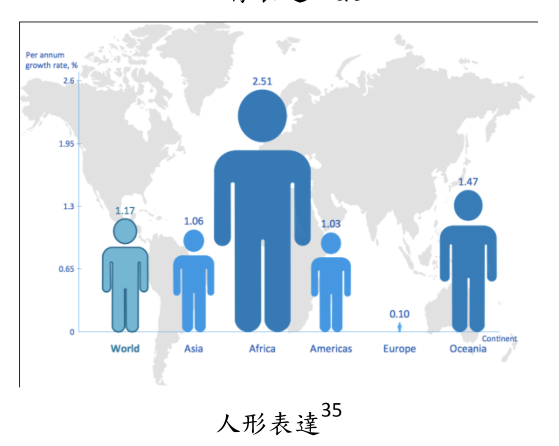
還有這些例子：



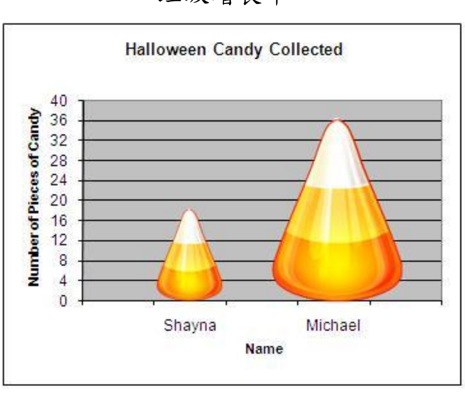
以人像表達人數[33]



垃圾增長率[34]

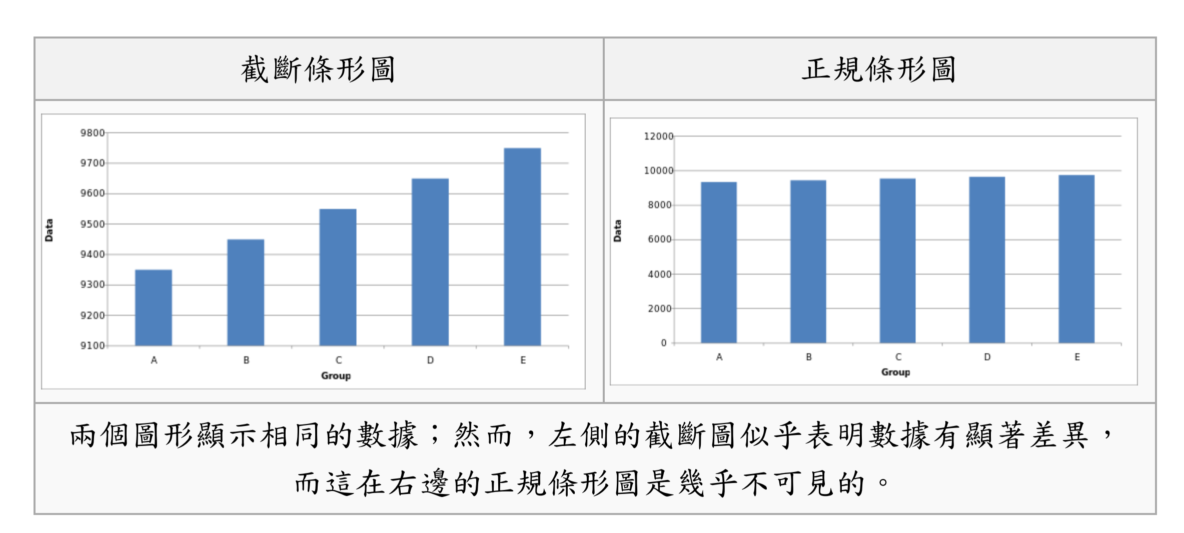


人形表達[35]



幾多倍？[36]

截斷圖形 truncated graph（也稱為撕裂圖 torn graph）的直軸（y 軸）不是從 0 開始，可用於顯示微小的變化或節省空間，但可能導致把少許變化錯認為重要變化的錯誤印象。如數值是在狹窄範圍，有些軟件（如 MS Excel）其默認功能會自動製作截斷圖形。



兩個圖形顯示相同的數據；然而，左側的截斷圖似乎表明數據有顯著差異，而這在右邊的正規條形圖是幾乎不可見的。

---

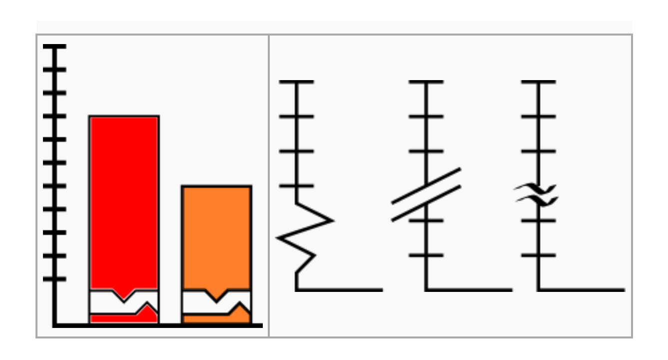
[33] http://www.timwallace.info/b/wp-content/uploads/2011/03/womendiagram.jpg

[34] http://yale.edu/ynhti/curriculum/images/2008/6/08.06.06.03.jpg

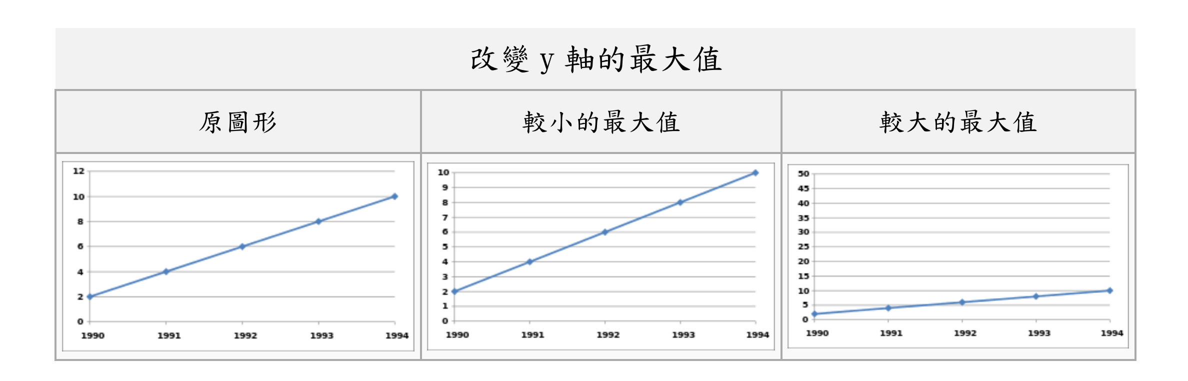
[35] http://www.conceptdraw.com/solution-park/resource/images/solutions/picture-graphs/GRAPHS-AND-CHARTS-Picture-graphs-Population-growth-by-continent-Sample.png

[36] http://yale.edu/ynhti/curriculum/images/2008/6/08.06.06.11.jpg

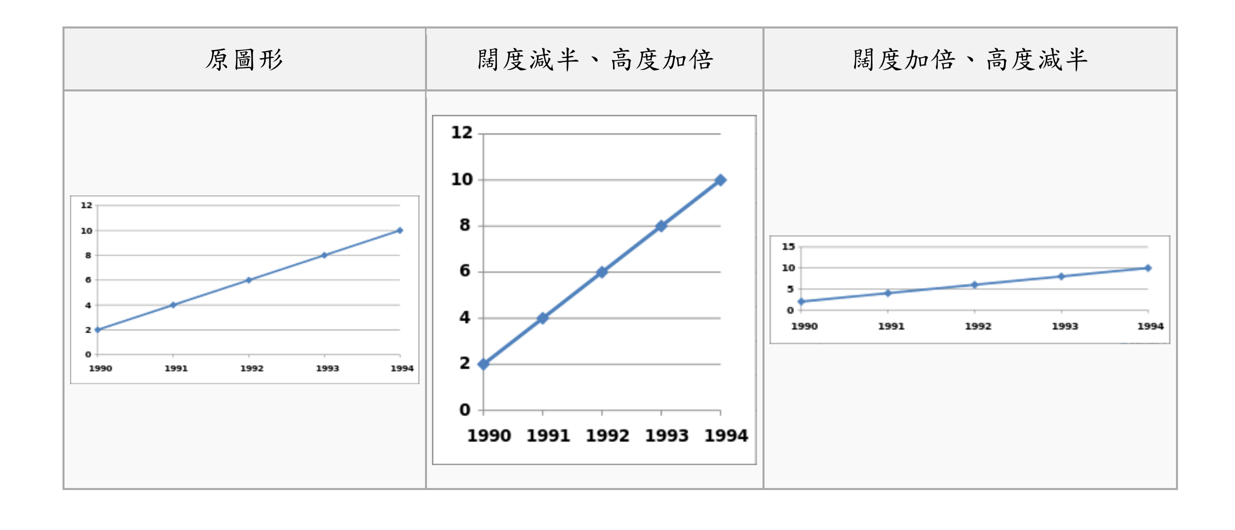
應適當提醒讀者直軸被截斷。



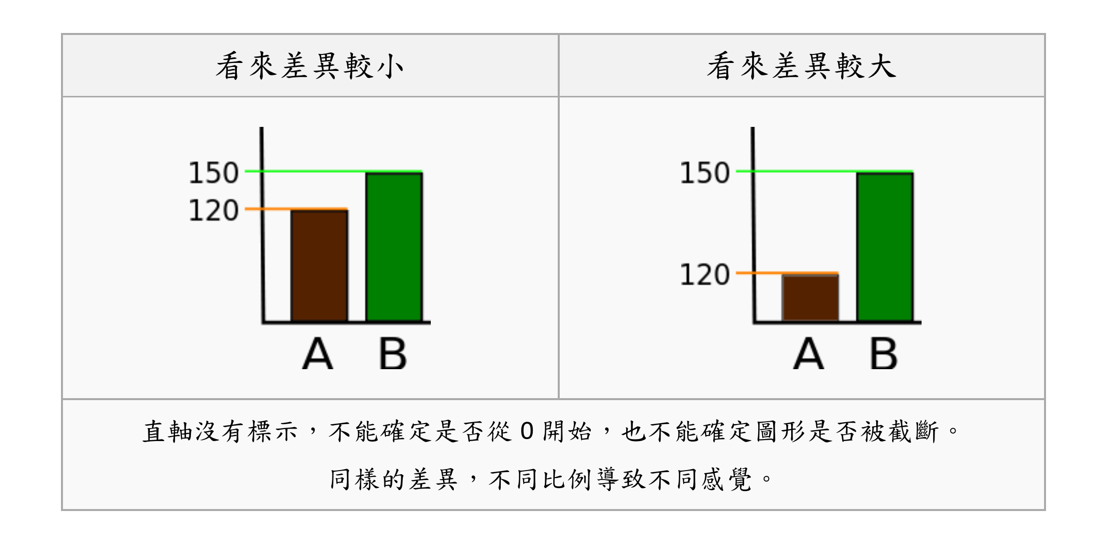
改變直軸的最大數值會導致不同的感覺。



改變圖形長闊比例會導致不同的感覺。



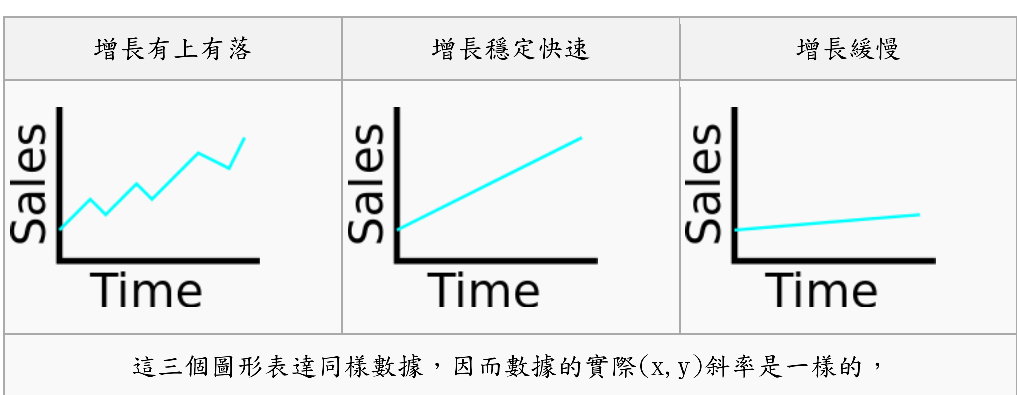
沒有比例的圖形往往用於誇大或減輕項目差異的感覺。



直軸沒有標示，不能確定是否從 0 開始，也不能確定圖形是否被截斷。

同樣的差異，不同比例導致不同感覺。

另一例子：

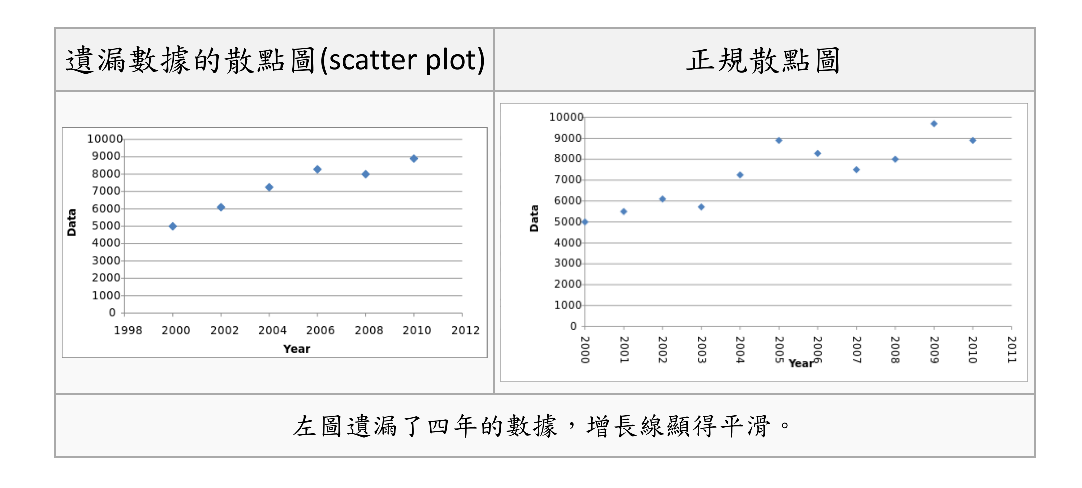


這三個圖形表達同樣數據，因而數據的實際(x, y)斜率是一樣的，

但因為各自的直軸比例不同（沒有標示），所以視覺上有不同的斜率。受操控的誤導。

數據遺漏

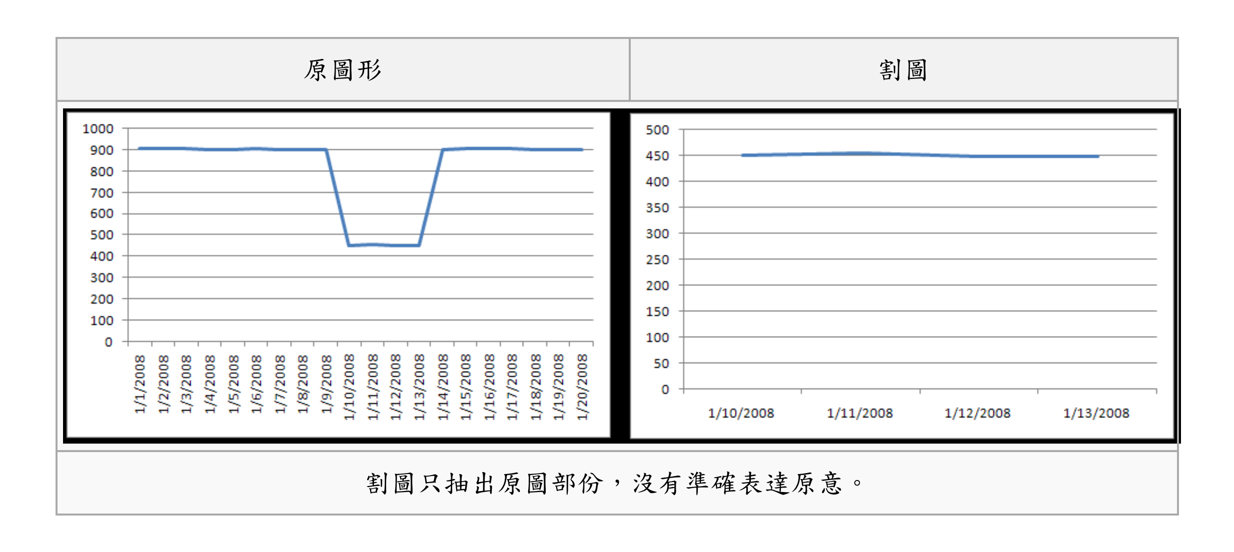
遺漏了數據的圖形就是誤導的圖形，不能從中得出正確結論。



遺漏數據的散點圖(scatter plot)	正規散點圖
左圖遺漏了四年的數據，增長線顯得平滑。	

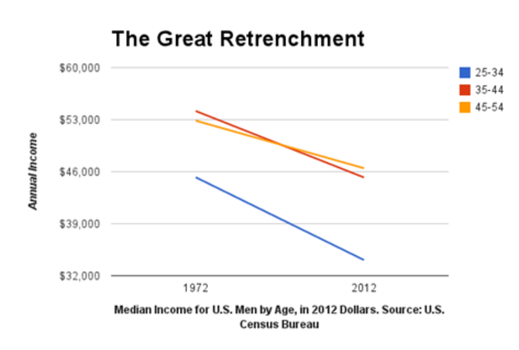
不正當的割圖

從其他圖形抽出部份為割圖，應保留（有時強調）原來的特徵。



原圖形	割圖
割圖只抽出原圖部份，沒有準確表達原意。	

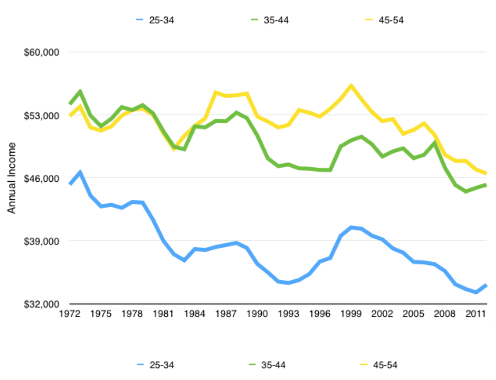
剪裁數據和扭曲圖形



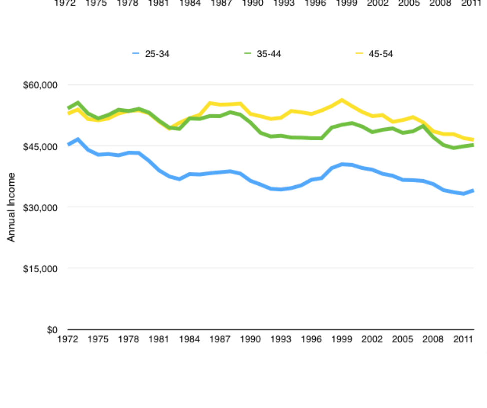
2013 年，彭博通訊社企業及市場編輯發表署名文章〈美國男士四十年來收入下降〉For U.S. Men, 40 Years of Falling Income，附上插圖說明三個年齡組群的美國男士的中位數收入下降，下降斜率頗為驚人。文章集中討論 1972 年和 2012 兩年的數據。

數據來自美國人口調查局，彭博是有聲譽的通訊社，作者不是初出茅廬的見習記者，報導應該是可信的吧？

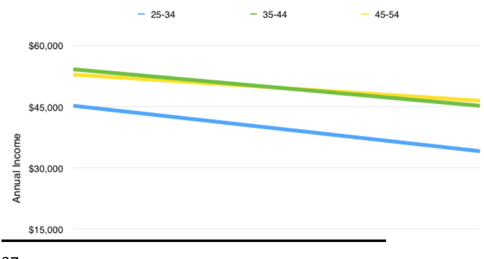
Eric Portelance[37]留意到這截斷圖（直軸不是從 0 開始）問題多多，於是深入研究相關數據，發現原作者只集中討論 1972 年和 2012 年的數據，似乎故意忽視了在這期間的多年數據。



重新製作的沒有截斷的連續圖給出不同年份的數據，得出不同印象。總體而言，中位數收入依然呈現下降趨勢，但斜率不是第一圖的劇烈。45-54 歲組群是相當穩定，直至 2000 年才有下降。



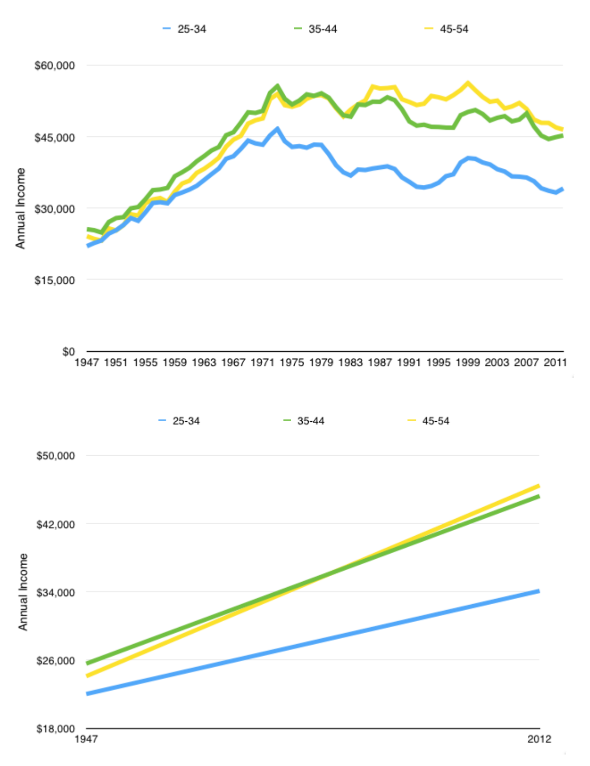
若是圖形沒有截斷，回歸正規從 0 開始，中位數下降的斜率可說是緩慢。



若原圖沒有截斷，中位數下降的斜率不是文章強調的「危險」。

[37] https://medium.com/p/c63780efa928

Portelance 進一步找出人口調查局的全部數據，發現彭博編輯「忽略了」1947 至 1972 年的趨勢。



1947 至 2011 年的全部數據得出不同的結論：收入持續上升，直至 1971 年見頂，之後有些年齡組群保持平穩，有些逐年下降。研究主題應該是「為何如此？」而不是「美國男士四十年來收入下降」。

如追隨彭博作者只選用兩年的數據作為起點和終點，不同的選擇(只選 1947 和 2012 年）得出完全不同的結論！

這是統計謊言的典型例子。

第七章　半吊子的數字

一名印度法官忠告熱心的年輕英國公務員：「當你年紀大一點，就不會熱衷於統計數據。印度非常熱衷於積累統計：收集，添加，提高至 n 次冪，取立方根，並準備精彩的圖形；但絕不能忘記的是這些數字每一個都是來自村長，他們喜歡說什麼數字就說什麼！」

如果不能證明你想證明的什麼，證明別的東西，假裝是同一東西。人們面對統計統計數據的衝擊時會發呆，幾乎不會注意到其中的差別。半吊子的數字是非常有用的手段。

藥廠不能證明新藥能治感冒，但可以大字發布實驗室報告：半公克新藥在試管內 11 秒殺死 31,108 枚病菌。要確保實驗室是有信譽或有令人印象深刻的名字。拍攝穿白袍的醫生拿著報告。

但不要提出幾個噱頭：在試管中有良好效用的藥劑可能不會在人的喉嚨有作用，不要說明殺死什麼病菌以免混淆。誰知道是什麼病菌引起感冒，特別病源可能不是病菌？事實上，沒有人知道試管中各種細菌和感冒有什麼關連，但人們不會深入理解，尤其是感冒病人。

也許，這例子太明顯了，人們多了對感冒的認識，雖然廣告頁面從來少不了這些聲東擊西的例子。

在種族歧視的年代，奉命調查以「證明」不是這回事，這是艱巨任務。你可以計劃一次民意調查，或更好的是委托有聲譽的機構調查；向有代表性的母體群發問：黑人的就業機會是否和白人一樣？每隔一段時間進行一次調查，最後得出趨勢的結論。

普林斯頓大學民意調查中心曾經調查這題目，發現得出的民意表裡不一。每位受訪者除了回答主題問題，還要回答其他問題以測試他是否歧視黑人。調查發現種族歧視觀念最嚴重的受訪者，對就業問題的答案往往是正面。同情黑人受訪者有三分之二認為黑人就業機會遜於白人；有種族歧視觀念的人有三分之二認為黑人就業機會不遜於白人。明顯這項調查對黑人公平就業機會說不清是什麼情況，反而揭露了人們看待種族的另一面。

因此，在種族歧視的年代，調查黑人的公平就業機會，會得出「黑人就業沒有問

題」的結論。情況越差，這些半吊子數據讓調查看來更好一些。

「執業醫生有 27%選擇金葉牌香煙，多於任何其他牌子。」暫由不論這說法是否虛假，只要問這說法有什麼問題。大多數人的反應可能是：「那又怎樣？」醫學界受到尊重，但醫生知道香煙品牌的訊息是否多於普通煙民？他們是否有特別知識選擇危害最小的香煙？當然他們不是這樣。然而，「執業醫生有 27%選擇金葉牌香煙」似乎意味著更多的什麼。

「實驗室試驗證明大力牌電動榨汁機功能提高 26%。」。這聽起來真不錯；直至真相揭露是大力牌電動榨汁機的功能是與老式手動搾汁機比較。大力牌電動榨汁機可能是市場上功能最差的，那個 26%數字是完全不相干。

不是只有廣告客戶玩弄數字，更多的是從數字中導出沒有關連的結論。一篇交通安全的文章報導：「晚間七時的交通意外是早上七時的四倍」，因此在早上開車更安全。數據沒有問題，但結論不可靠。晚上的交通比早上繁忙，所以較多意外，與文章的結論沒有關係。

如果沒有留意這些數字是半吊子的數據，你可以被任何交通工具事故的統計數據嚇得半死。

相比 1910 年，更多人死於飛機意外。現代的飛機是否更危險？廢話。現在的飛機乘客是以前的數百倍，僅此而已。

「據報導，去年的鐵路意外死亡人數為 4,712 人。」這很嚇人。真相是有一半死亡人數是因為汽車司機闖紅燈，在道口與火車相撞，其餘大部份是跳車的霸王乘客，只有 132 人是火車乘客。甚至這數字也沒有很大比較意義，除非這連接到總乘客里程。

知道火車，飛機或汽車去年的意外傷亡數字，也要同時知道每百萬乘客一公里數字，才可以知道風險比率。

聲東擊西有很多法寶，一般手法是並列兩種看來相關或相似，但其實沒有關連的項目。某企業與工會的關係惡劣，人事部經理受命「調查」員工對工會的投訴，必然可以找到一些相關投訴，理直氣壯聲稱「員工有 78%反對工會」；實情只是搜集一些不經分類的投訴和埋怨，彙集為另一套數據。這沒有證明什麼，但似乎是完成了調查。

當然，這是雙面刃；工會也可以隨時「調查」，「證明」員工對企業的諸多不滿。

企業的財務報告多的是這些半吊子數字。留意出乎意料的龐大利潤和隱藏在某他名目的利潤。汽車工人工會有這樣的報導：

「公司公報去年利潤三千五百萬元，佔銷售額的 1.5%」，少得可憐。換一個三毛錢的燈泡已耗上二十元銷售額。員工甚至想到要節省用紙。公報的利潤當然不是全部利潤，其餘的隱藏在折舊，特別折舊和儲備。

同樣要留意百分比。通用汽車公報本年九個月的稅後銷售利潤增加 125%，投資部門盈利增加 448%。這究竟是好是壞？視乎你的觀點。

同樣，讀者來函為 A&P 商店辯護：「商店每千美元銷售額只賺了十元，不應被譴責為奸商。」咋聽之下，這樣的利潤確實微不足道；住房抵押貸款和銀行貸款的息率在 6%之上。A&P 公司結束超市業務，把資金存入銀行賺取利息豈不是更有生意頭腦？

心法在於投資年回報率不是等於銷售總額的利潤。正如另一位讀者投函解釋：「如每天早上以$0.99 買貨，當天以$1 價格售出，利潤只有 1%，但全年 365 天的投資盈利是 365%。」

任何數字都有許多表達的方式。例如，可稱之為銷售回報率 1%，投資回報 15%，一千萬美元的利潤，利潤比去年增加 40%，或比去年下降 60%，方法是選擇一個最適合當前目的的數字，希望沒有幾個人會理解這是如何不完善反映了情況。

不是所有半吊子數字是故意欺騙的產品。許多統計數據，包括對大家非常重要的醫療數據，是因為源頭失真而被扭曲。一些微妙事項如墮胎，婚外生育和梅毒都有驚人的矛盾數據。美國最近公佈的流感和肺炎數字，奇怪的結論是這些疾病幾乎都局限在南部三個州，佔報告病例約 80%。實情是這三個州依然把流感和肺炎列為必須申報的病例，其他州已經停止申報。

一些關於瘧疾的數字沒有意義。1940 年前，美國南部每年有數十萬例，現在只有極少數，似乎短短幾年內有極大改進。實情是現在只呈報確診為瘧疾的病例，而之前是包括了南方人慣稱的感冒或發冷。

1898 年的美西戰爭，海軍死亡率是 9‰，同一時期的紐約市平民死亡率是 16‰。海軍徵兵人員後來用這些數字來宣傳在美國當海軍更安全。假設這些數字是準確的，看看這兩個數字為何幾乎毫無意義。兩個組群沒有可比性。美海軍主要身體健康的年輕人；紐約市平民包括嬰幼兒，老人和病人，他們全都有較高的死亡率。

兩個數字不能證明符合海軍標準的士兵活得更長壽，但也不能反證。

在發明脊髓灰質炎疫苗之前，沮喪的消息是小兒痲痺症是史上最嚴重，當年比以往任何時候都更多病例。

專家檢視這些數字，發現幾件令人鼓舞的事情。其中之一是當年的小兒數目是破紀錄的數字，如發病率不變，病例數字也會水漲船高。另一發展是父母更多認識脊髓灰質炎，即使輕症病例更願意求醫就診。最後是有了財政誘因：有更多的小兒痲痺症保險和慈善組織的更多援助。所有這一切令人懷疑小兒痲痺症達到新高的說法，後來的死亡總人數證實了懷疑是合理的。

值得留意的事實是死亡率或死亡人數往往比發病率或發病人數是更好的衡量 - 僅僅是因為報告和記錄死亡率或死亡人數是較為盡心和準確。

美國每四年就有一次半吊子數字的熱潮。數字沒有週期，而是四年一度的選舉來了。共和黨在 1948 年 10 月發表的競選聲明完全是建立在似乎是互相關連但原來互不相關的數字：

> 1942 年，當 Dewey 當選州長時，一些地區老師的的最低工資低至每年$900。今天，紐約州學校的老師享有世上最高的薪水。Dewey 州長接納他委任的委員會調查結果，在 1947 年提取部份盈餘即時增加教師薪金。因此，紐約市教師的最低薪金是$2,500-5,325。

完全可能 Dewey 先生是教師之友，但數字不是這樣說話。這是比較「之前」和「之後」的老把戲，從$900 急增至$2,500-5,325 聽起來是極大改進，但沒有說明$900 是農村地區教師的最低工資，而$2,500-5,325 只是紐約市的範圍。Dewey 州長可能改善了教師的待遇，也可能沒有。

之前和之後的比較照片是雜誌和廣告的熟悉特技。拍攝兩次，告訴你新油漆塗層可以做到什麼區別。在兩次攝影之間，客廳已經添加新傢具，有時「之前」的照片只是很小，光線不好的黑白照，「之後」版本是全彩色大照片。比對照片顯示模特兒用護髮素的前後對比：天哪，她確實好看得多，但仔細檢查會發現大部分的變化是因為她的微笑，光亮頭髮。是攝影師的功勞，不是護髮素。

補充材料



2007 年，英國的廣告聲稱：「多於 80%牙醫推薦高露潔牙膏」。一般人從廣告得出的印象是 80%牙醫推薦高露潔牙膏，餘下的 20%推薦其他牌子。

英國廣告標準局介入調查，發現數據來自高露潔贊助的市場調查（但沒有公佈），而且受訪牙醫可以推薦多款牙膏，不是只選一項。調查數據顯示至少有另一牌子和高露潔的得分不分上下。

英國廣告標準局下令禁制廣告。[38]



2009-10 年，體育用品公司 Reebok 聲稱 EasyTone 和 RunTone 跑步鞋經實驗室測試，「證明只需穿上跑步鞋走路，比一般跑步鞋有助強化腿筋和小腿 11%，臀部肌肉更高達 28%！」

美國聯邦貿易委員會調查發現這完全沒有科學根據，被判罰款二千五百萬美元。[39]



〔台灣〕行政院公平交易委員會委員會 27 日決議，台灣莊臣公司在贈品包裝上登載「近 90%消費者選擇植物歐護」，商品品質及內容為虛偽不實及引人錯誤，違反公平交易法規定，處新台幣 100 萬元罰鍰。

中央社報導，公平會表示，台灣莊臣依據博輿市場研究顧問於 2006 年 7 月間進行的市場問卷調查，在其贈品包裝廣告上宣稱，近九成消費者「選擇」植物歐護。

公平會指出，但經調查，該問卷其實是將莊臣的歐護植物防蚊液與另一品牌防蚊液，進行清爽不油膩偏好的比較，而非購買的比較，廣告卻未註解「九成」的比較基礎，恐致消費者誤導。

[38]資料來源：http://www.telegraph.co.uk/news/uknews/1539715/Colgate-gets-the-brush-off-for-misleading-ads.html
[39]資料來源：http://www.investopedia.com/financial-edge/0612/4-examples-of-misleading-health-ads.aspx

公平會表示，此外，該問卷調查以隨機抽樣方式進行，就 100 位受試者現場使用兩種產品後調查，姑且不論樣本數是否足以支持該廣告宣稱內容，廣告宣稱「近九成消費者選擇歐護」，顯然與問卷調查結果有別，因此認定廣告不實。[40]



Centrum 在 1997 年的廣告聲稱「十個美國人有九個未能從食物攝取所需的營養素，缺少了重要的維生素和礦物質。」該聲明引用 1976 至 1980 年間進行的一項調查，發現在調查當天，受訪者只有 9%記得要進食水果和蔬菜的每日推薦量，因此得出結論高達 91%的美國人缺少維生素（可能包括你！）。

這說法問題多多：（一）這不能證明那些人缺少維生素；事實上，他們可能在前一天已進食足夠數量的水果和蔬菜；（二）只是一天的飲食不足以計量整體飲食習慣。食物攝入量應以幾星期計算；（三）即使攝入數量低於推薦量也可以有充足營養。[41]



Vioxx 是一種非甾體抗炎藥，類似阿司匹林或布洛芬。Merck 藥廠的直銷廣告耗資億萬美元（2000 年花費了 1.6 億美元）。該藥物於 1999 年被 FDA 批准，直至 2004 年才停用。這是源於一宗法律訴訟聲稱該藥物引起 23,800 宗心血管病例（包括心臟病發作），跟進研究發現服用 Vioxx 的患者其心血管病例統計上顯著高於安慰劑患者。

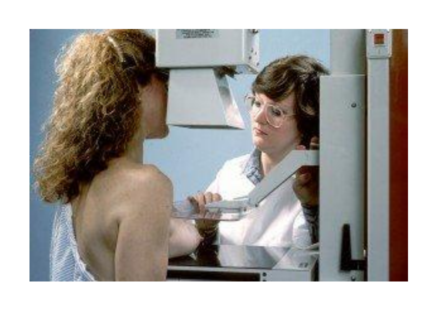
這種不安全藥物如何得到 FDA 批准推出市場。事因原有研究發表時，藥廠排除了三宗心肌梗塞的病例，從而改變了統計顯著性。可以想象藥廠僱用的科學在重重壓力下「忘記」這三個病例，或是他們不理解統計顯著性的意義。[42]



1995 年，英國藥物安全委員會向十九萬名醫護人員發出警告：第三代口服避孕丸增加了在腿部或肺部形成血

[40] http://dasanlin888.pixnet.net/blog/post/34467926
[41] 資料來源：http://www.statisticshowto.com/misleading-statistics-examples/
[42] http://www.statisticshowto.com/how-significant-is-significant-the-vioxx-scandal/

塊，有潛在的雙倍致命風險。這警告導致在 1996 年有一萬三千宗墮胎手術。所謂「潛在的雙倍致命風險」原來是基於以下的數據：每十萬名服用第二代口服避孕丸的婦女有十五人患上可致命的血塊；服用第三代口服避孕丸的則增至二十五人。作為參照，沒有服用避孕丸的婦女每十萬人有五宗病例。是的，風險是增加了，但比懷孕的風險要小得多，不值得那麼令人震驚。[43]



統計師被醫生告知她的乳房 X 線檢查呈陽性反應，她詢問醫生她患癌的機率是多少？。醫生給出令人震驚的答案：80%。她遍查文獻，找到正確答案是 10%，更令她震驚的是許多醫生給出不同答案：20%醫生回答 10%、20%醫生回答 1%、60 %醫生回答 81 或 90%。

不是醫生看不懂數字，而是有太多研究報告被斷章取義，渲染誇大。[44]

---

[43] http://news.bbc.co.uk/2/hi/health/313848.stm

[44] http://www.statisticshowto.com/even-physicians-dont-understand-statistics/

第八章　「後此謬誤[45]」又來了

要估算荷蘭或丹麥的家庭生了多少孩子，你可以亂猜，或是計數他們房子屋頂的鸛巢。[46]

統計術語描述鸛和新生兒兩者之間有「正相關關係」，有 A 就有 B。

這個古老神話實際說明更有價值的意義：容易記住和提醒我們兩個因素之間的關聯不足以證明在前的 A 引起了其後的 B。

在鸛和嬰兒的例子，很容易找到與兩者相關的第三個因素：大家庭住在大房子，大房子有更多煙囪讓鸛鳥築巢。

但在其他情況，不總是那麼容易發現因果關係的假設缺陷，尤其是流行偏見認為這是有特別意義。

有人研究和證實煙民的大學成績是低於非吸煙者。很多人很高興，這說法流傳到現在。這樣看來，要有好成績是在於放棄吸煙；再進一步的結論是吸煙讓人變蠢。

我相信這項研究是正確完成：有誠實和精心挑選的足夠樣本，相關性高等等。

其中的謬誤頗為古老，經常出現在統計材料，躲在可觀的數字之下。謬誤就是：因為先有 A，後有 B，所以 A 導致 B。既然吸煙和學業不走在一起，因此吸煙導致學業不佳。但也可以倒轉來說：學生成績不佳驅使他吸煙草，但不酗酒；這結論也可以證明是對的，也得到證據的支持。但這不能滿足宣傳手法。

更好的結論是兩者沒有關連，兩者都是第三因素的產物。是否喜歡交際的學生較少時間看書而多抽煙？或者之前某人證實外向性格與成績低落之間有相關，這關係比成績與智力之間關係更為明顯？也許外向性格比內向的人更多抽煙。問題的關鍵是有很多合理解釋，很難只是堅持己見只挑選一個。但很多人是這樣。

為了避免掉落「後此謬誤」的謬論作出錯誤判斷，你需要仔細檢查任何關乎「彼此相關」的陳述。這種謬誤有幾種類型。

---

45 Post Hoc 一個事件發生在另一事件之前，並不一定是後者的原因，也譯為「事後謬誤」。

46 圖片取自 http://www.todayifoundout.com/wp-content/uploads/2013/05/stork-340x400.jpg。歐洲民間傳說鸛是送子鳥。

一種是偶然產生的相關性。你可搜集一組數字來證明一些不太可能的事情；但如再試一次，可能無法證明。一如「牙膏防止蛀牙」的廣告，你只需扔掉不想要的結果，廣泛發佈那些合心意的結果。如只是小樣本，很有可能發現你想得到一對一事件之間的一些實質性關聯。

常見的一種共變是其中的關係是真實的，但不可能確定那個變量是「因」，那個是「果」。在某些情況下，因果關係可能會時不時改變從屬地位，或兩者可能同時是「因」也同時是「果」。人們的收入和持有股票之間的相關性可能是這樣。有更多錢就多買股票；有更多股票，收入越多；說不準是那一個導致另一個。

也許最棘手的是變量互不影響，但有真正的相關性。這方面有頗多研究，例如煙民的學業成績差勁；有太多醫學統計雖然證實相關關係是真實的，但這「因 A 而 B」的關係只是猜測而矣。作為廢話或偽相關的統計例子，有人興高采烈地指出：馬薩諸塞州長老會牧師的薪金和古巴甜酒價格有密切關係。

何者為「因」？何者為「果」？換句話說，長老是否受益於或支持甜酒貿易？這太牽強了，明顯是荒謬之言。緊記世事多的是「後此謬誤」，只是更為微妙隱蔽。長老和甜酒的例子很容易看到這兩個數字齊齊增長，是因為第三因素的影響：世上萬物的價格都在增長。

〔歐洲〕人們提到六月的自殺率最高，也提到最多人在六月結婚。是否自殺驅使較多人結婚？或是較多求婚不遂的人自殺？稍微更有說服力（但同樣未經證實）的解釋是在整個冬天舔著抑鬱傷口的人本以為到了春天會雨過天晴，可是六月來了，他仍然感到絕望，…。

要注意的另一個結論：推斷得出的相關性已超越引以為證的數據。很容易表明多雨水，玉米和農作物生長得更高更好。似乎雨水是好事。但連綿數月的強降水會損壞甚至破壞農作物。正相關關係只能維持到某一點，然後好事變壞事。超過一定的雨量，下雨越多，玉米收成越少。

當然，「相關性」的傾向經常不是被描述為一對一的理想關係。高個子男生的體重超過矮子男生，這是正相關關係。但是可以很容易找到一個六英尺的高個子體重及不上五英尺的矮子，所以相關性是小於 1。負相關簡單說明「此消彼長」：變量 A 增加，變量 B 會下降。在物理學這是「反比」：燈泡的光線越遠越弱。這些物理關係往往有完美的相關性，但是企業或社會學或醫學數字很少是如此整齊。即使學歷一般與收入成正比，但往往有許多反證。請記住，相關性可能是真實和基於真實因果關係，但如在單一事件中確定任何行動，可能是幾乎一文不值。

有無數研究證實大專以上學歷與未來收入掛鉤，大學派發無數小冊子吸引學生。我不否定這意圖，我贊成教育，特別是課程包括《統計學入門》。這些數字已經明確證明擁有大學學位的人賺更多。當然，有很多例外情況，但趨勢是強勁和明確的。

唯一的錯誤是有人利用這些數字和事實得出完全沒有根據的結論。這是後此謬誤的最佳例子。有人認為這些數字表明：如果你上大學，在這三、四年間你可能賺到的收入是高於以其他方式消磨這三、四年。這種沒有根據的結論其依據是基於同樣毫無根據的假設：因為曾受大學教育的人賺更多錢，是因為他們上過大學。其實我們不肯定知道：這些人即使沒有上大學，可能都會賺更多。一些事實強烈表明正是如此。大學學生有兩個群組多得不成比例：富家子弟和聰明學子。聰明的人即使沒有上大學，可能已經有很好的賺錢能力。談到富家子弟…錢生錢有多種方式。無論是否上大學，富家子弟很少落在低入息階層。

銷量龐大的星期日報刊有以下這段對話，也許你會覺得有趣，因為同一作家有另一篇文章〈流行觀念：對或錯〉。

> 問：上大學對你終生不結婚的機會有什麼影響？
> 答：如果是女生，一生老處女的機會挺高。男生剛好相反，很少終生不娶。

美國康奈爾大學調查 1,500 名典型的中年大學畢業生，發現男生有 93%已成婚（相對於一般人口只有 83%）。但中年女性畢業生只有 65%結了婚。大學畢業生中的老處女是一般人口終生不嫁婦女的三倍。

十七歲的小美看到報導，知道如果她去上大學，婚姻大事的前景很不樂觀。而且統計資料的來源頗有聲譽。是的，報導有引用康奈爾大學的統計數據，但結論不是倉促讀者所認為是來自校方的。

這又是案例：利用真正的相關性強加諸未經證實的因果關係。也許這一切是倒過來說。即使這些女生沒有上過大學，仍然會終生不嫁，比例甚至可能高於大學女生。如果這說法的可能性並不優於作家堅持的結論，這也許也是猜測而矣。

事實上，有證據表明有終生不嫁傾向的女士更有可能上大學。金賽性學博士似乎找到了性慾和教育有一定相關性，而性狀可能在大學預科年齡期已形成。這更令人質疑上大學會影響人們結婚的說法。

所以，小美注意：這是未必如此。

醫學文章曾經提出嚴重警告，指出喝牛奶的人患癌的機會增高。在美國新英格蘭，明尼蘇達州，威斯康星州和瑞士這些大量生產和飲用牛奶的地方，癌症似乎變得普遍，而在牛奶稀缺的亞洲國家斯理蘭卡罕見癌症。文章也指出美國南方各州少喝牛奶，癌症病例也較少。此外，有人指出經常喝牛奶的英國婦女患上某些類型癌症是少喝牛奶的日本婦女的十八倍。

只要稍為深入研究這些數字就可以得出不同解釋。癌症主要是中年或以後的疾病。瑞士和前文提到的國家同樣的是國民長壽。在那項英日婦女研究，英國婦女比日本婦女平均年長十二年。

Helen M. Walker 教授提出證明，解釋有趣但愚蠢的說法；證明假設每當兩件事情一起變化必然有因果關係的謬誤。調查婦女的年齡和某些物理特徵之間的關係，可以計算步行時腳的角度，會發現老年婦女的角度往往較大。可能即時反應這反映因為腳的角度加大，所以她們長老了。人人都看出這是荒謬的解釋。似乎是年齡增長導致腳的角度增大；大多數婦女長老了，腳的角度加大。

任何這樣的結論很可能是虛假和必然是不合情理。要適當得出正確結論，研究應在一段時間內觀察同一婦女或類似組群。這會消除一個可能的因素：老年婦女成長時，被教導走路時腳要朝外，而現在的年輕少女被教導這樣的姿勢不正確。

如有人（通常是有利害相關的人）對某項相關關係大做文章，首先看看這是否這類型的關係：產生於事件流程，時間趨勢。我們這時代很容易發掘到任何兩項事物的正相關關係：大學學生人數，精神病人數目，香煙消耗量，心臟病數字，使用 X 光機次數，加州學校教師的薪俸等等。認為其中一些事物是另一些事物的「因」顯然是愚蠢無理。但太陽之下無新事，每天都有人提出。

以統計學方法和迷惑的數字和小數點來闡釋因果關係，只是比迷信好一點，但往往比誤導更嚴重。新赫布里特群島的島民一直相信體蝨是健康良好的表徵。他們觀察了幾百年，目睹身體健康的人通常有體蝨，而生病的人往往沒有。觀察本身是準確和有見識；歷久以來，這些非正式的觀察往往都是。從證據中得到這些原始結論：體蝨讓人健康，人人都應該有體蝨。對此，我們很難有什麼說法。

正如上文指出，在統計磨房處理比這還要稀少的數據，直至常識的目光再也無法穿透，已經為醫療界和許多雜誌和專業醫學期刊賺錢不少。精明觀察者終於弄清楚新赫布里特群島的現象。事實證明，幾乎每個島民大部分時間都有體蝨；可說是正常狀態。然而，當病人發熱（很可能是由那些體蝨傳染），病人體溫變得太熱，體蝨離開這不再舒適的居所。這案例的因果完全混淆、扭曲、扭轉和混在一

起。

補充材料

錯誤的因果關係

當統計測試展示 A 和 B 之間的關係，通常有五種可能性：

1. A（因）導致 B（果）。
2. B（因）導致 A（果）。
3. A 和 B（因）互相導致對方出現（果）。
4. A 和 B（因）一起導致 C（果）。
5. 觀察得的關係純屬偶然（沒有因果關係）。

第五個可能性可透過統計測試來量化，計算出來的機率與前四個可能關係發生的機率一樣大，但事實上應變量之間是沒有關係。

如調查發現沙灘泳客購買雪糕的人數與遇溺人數有相同趨向，沒有人會斷言雪糕導致遇溺，因為這是明顯地無關。遇溺和購買雪糕的人數明顯與第三個因素（沙灘上的人數）相關。

但這謬誤的例子不是笑話：例子是「接觸化學品 X 會導致癌症」的諸多報導。把「接觸化學品 X 的人數」代替「購買雪糕的人數」；把「患上癌症的人數」代替「遇溺的人數」。在這情況下，即使兩者沒有真正的因果關係，但統計上依然有關聯。例如，如某地方有「危險」（即使並不危險）的化工廠，中產家庭因恐懼而遷離，誘使更多低收入家庭搬到該地。然後發現低收入家庭患上癌症的數字上升，於是推論化工廠是元凶；其實這可能是基於較差的膳食和生活環境或是較低檔次的醫療服務。

第九章　統計誤世

通過使用統計材料以誤導他人，可稱為統計操控，或是「統計誤世[47]」。

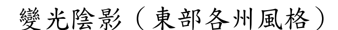
本書的書名和一些內文似乎暗示所有這些操作都是意圖欺騙的產物。美國統計協會的分會會長曾為此斥罵我。他說：大多數不是欺騙，而是無能。他的說話有意思，但我不能肯定統計學家認為那一項批評更為不恭敬。可能更重要的是要記住：扭曲統計數據及其操作並不總是專業統計人員所為。統計學家的成果被推銷員，公關專家，記者，或廣告文案扭曲，誇張，過度簡化，或通過選擇扭捏。

但無論在任何情況下誰是有罪的一方，很難說這是無心之失。雜誌和報紙經常誇大炒作虛假的圖表，很少減斤扣兩。在我的經驗，業界提出的統計參數很少報大報喜，往往是表達差於數據。另一方面，少見工會聘請無能的統計人員做出數據差於表達的統計。

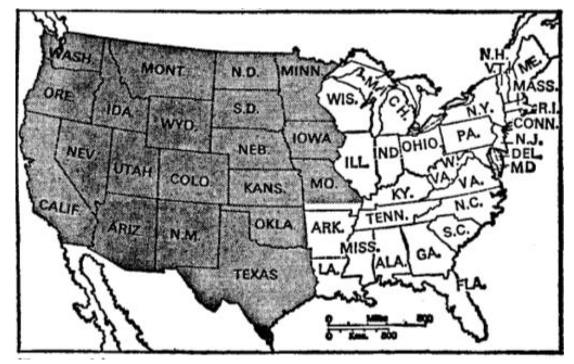
只要這些錯誤是一面倒，很難歸結於笨拙或意外。

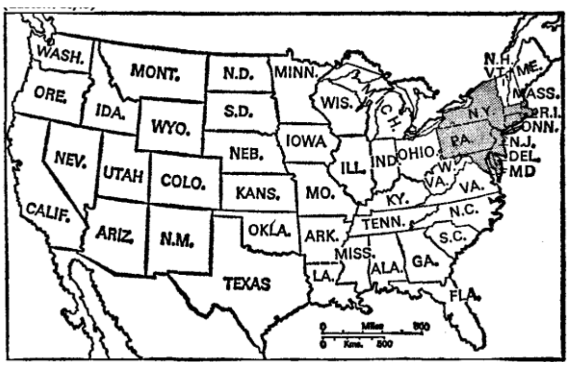
歪曲統計數據巧妙手法是利用地圖。地圖隱含許多變量，其中事實可以被掩飾，關係被扭曲。我最喜歡的「變光陰影[48]」獎杯頒發給不久前波士頓第一國民銀行發表和轉載極廣，包括所謂納稅人群體，報紙和《新聞周刊》。

變光陰影（西部各州風格）



變光陰影（東部各州風格）





為了表示我沒有作弊，地圖加了 MD, DEL 和 RI。

該圖顯示目前聯邦政府拿走和花費的美國收入部份，利用有色部份表示密西西比河以西各州（除了路易斯安那州，阿肯色州和密蘇里州部分），其聯邦政府支出

[47] statisticulation

[48] The Darkening Shadow

已等於各州國民的總收入。

欺騙謊言在於選擇地廣人稀的各州，其收入相對較少。以同樣的誠信（和同樣的不誠信），繪圖者可能已開始在紐約或新英格蘭著色，得出極為更小但更令人印象深刻的陰影。使用相同數據，他可以給出產生完全不同印象的地圖，但沒有人有興趣發表。至少，我不知道有任何強大群體有感興趣發表偏少的公共開支。

如果繪圖者目標只是傳達訊息，很容易做到。他可以選擇一組中間狀態的州份，其總面積與總收入佔國民收入比例相同。

這張地圖公然誤導，不是宣傳的新把戲，而是經典手法。同一家銀行不久前公佈顯示聯邦政府在 1929 年和 1937 年開支的地圖版本，很快被輯錄為「可怕插圖」歪曲事實的例子。這間銀行依然故我發表繪圖，而更有見識的《新聞周刊》和其他人一直照搬可也，沒有警告也沒有道歉。

如果你認為現在有通貨膨脹，看看這個。有一段時間，美國人口普查局想出了在年報陳述「平均家庭收入為$3,100」。但同時報章又報導 Russell Sage 基金會給出的同樣數據是可觀的$5,004。也許你高興知道大家生活得不錯，但也可能感受到這數字與你觀察所得不符。也許你認識的人不是基金會認識的群組。

人口普查局和基金會的數字怎會如此不同？普查局是說「中位數」，也是應該如此；但即使基金會是說「平均數」，差別也不應該如此巨大。基金會解釋數據來自把美國人民個人總收入除以 149,000,000，得出人均$1,251；四口之家即共有收入$5,004。

這樣奇怪的統計操控有兩方面的誇大：（一）使用「平均數」而不是較小和更多訊息的「中位數」（上文有討論）；（二）假設家庭收入是家人數目成正比。我有四個孩子，也希望事情是這樣，但事實不是。四人家庭的收入絕對不是兩人家庭的兩倍。

公平地說，基金會的統計學家可能不是存心欺騙，應該說他是想表達人們捐獻而不是受惠的意思。有趣的家庭收入數字只是副產品，但這欺騙行為已廣泛傳播；這是不能輕信平均數的最好例子。

表面精確會賦予最聲名狼藉的統計數據看來有斤兩。考慮小數點的例子。調查一百人昨晚睡了多少小時，比如說得出總數為 7,831 小時。首先，任何這樣的數據遠遠不可能精確。大多數人的的猜測有十五分鐘或更長時間的錯誤，而且不能保證這些錯誤〔在數據集〕會自我平衡。有人失眠五晚，只記得折騰了半晚。無論

如何，調查算出各人的平均睡眠時間為 7.831 小時，聽來你是知道自己在做什麼。如果發表的數字是 7.8（或近乎 8）小時，這不是什麼驚人的吧。這是拙劣的接近數值，比幾乎任何人的隨意猜測都沒有什麼啟發性。

馬克思以同樣手法製造精密的虛假氛圍。他要計算工廠的「剩餘價值率[49]」，開始彙集一些假設、猜測和整數：「假設廢品為 6%…。成本為整數 342 英鎊。有一萬個紗錠…假設成本為 1 英鎊。折舊率假設為 10%。假設工廠租金為 300 英鎊。這些數據是由一位曼徹斯特市紡紗工人提供，可以信賴。」馬克思利用這些近似數值算出剩餘價值率是 6%。[50]

百分比是製造混亂的沃土。一如令人印象深刻的小數點，百分比為不精確數據罩上精密的光環。美國勞工部曾表示華盛頓特區的兼職家庭在指定月份領取的交通津貼，有 49%是每星期 18 美元。細查之下，這個百分比原來出自兩個只有四十一項優惠的案例。基於少數案例的任何百分比都可能誤導；直接給出數字更能提供更多訊息。如百分比帶上小數點，小心欺詐。

「現在購買聖誕禮物，節省 100%！」。這廣告聽來像是聖誕老人自掏腰包，但只是製造混亂。原來是減價 50%。節省 100%是指新價格的 100%；這是事實，但不是廣告吹噓的事實。

標準石油公司的文獻走得更遠：「割價 14～220%」。這似乎要求賣方支付買方一筆可觀費用去拉走油膩膩的東西。

某公司宣布貨品銷售獲利 3,800%，算自成本 1.75 元和售價 40 元。計算利潤百分比有多種方法（必須說明）。如果以成本計算，利潤率是 2,185%；以售價計算是 95-6%。這間公司發明了新方法，得出了誇張的數字；而這似乎常常發生。

甚至紐約時報轉載美聯社報導時，也犯了「移動基數[51]」的錯誤：「經濟蕭條今天狠狠地打了工人一記重拳。印第安納波利斯建築貿易工會屬下的管道工，泥水匠，木匠和其他工獲得工資增加 5%。這只是他們去年削減工資 20 %的四分之一。」

表面看來這算法很合理；但跌幅是基於一個基數（工人之前的工資），而今年的加薪是基於另一個較小的基數（現有薪酬水平）。

小小心算即可指出以上是統計誤算。為簡單起見，假定原來工資是每小時$1，削減 20%即是下跌到$0.8。$0.8 增加 5%即為$0.04，這是削減額的 1/5，不是 1/4。

---

[49] rate of surplus-value
[50] 看不清原文的計算方式，籠統譯之。
[51] Shifting Base

一如許多誠實謊言，這篇報導誇大了一個本來很好的故事。

這一切說明：要抵消減薪 50%，下一次加薪必須爭取 100%。

「轉移基數」做成許多折扣的錯覺。「五折再八折」不是原價的三折，而是四折，因為「八折」是以較小的「五折價」為基數。

一種裝模作樣的欺騙手法是把不對號但似乎相關的東西相加。一代又一代頑童都用這一套證明他們不用上學。

你可能還記得吧。一年 365 天，減去在床上度過的 122 天（三分之一），再減去飲食時間 45 天（每天三小時）。剩餘的 198 天要扣了 90 天暑假和其他假期 21 天。剩下來的日子甚至不夠分配給週未。

你可能認為大企業不會利用這古老和明顯的伎倆，但美國汽車工會堅持汽車企業依然用這一套來對付他們。

每一次罷工期間都會出現這謊言：汽車企業聲稱罷工每天的損失是若干百萬美元。這數字來自如罷工工人全力工作會製造出來的汽車，加上供應商的損失。一切可能的被加進來，包括銷售商的損失。

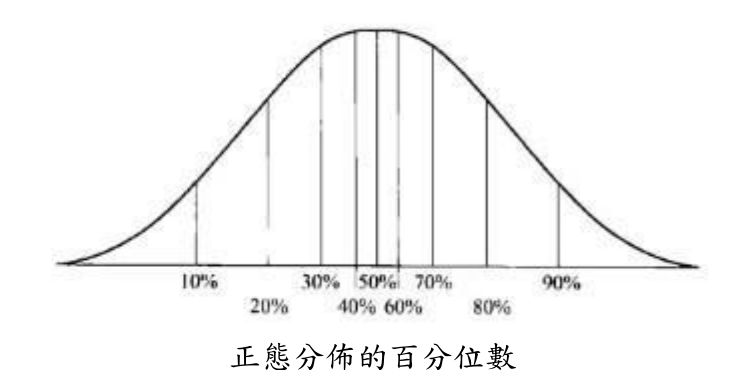
同樣奇怪的概念是百分比可以自由加在一起。《紐約時報》書評版這樣說：書價和作者收入之間的差距越來越大，是由於生產和材料成本大幅上升。在過去十年，廠房及製造費用上升多達 10-12%，材料上升 6-9%，銷售及廣告開支向上攀升超出 10%。只是一間出版社，這些林林總總加起來至少有 33%；較小規模的出版社幾近 40%。

其實，如果每個成本項目上漲約 10%，總成本必然也以 10%同樣比重攀升。把各項成本的增加疊加起來，是鬼話連篇。今天你買了二十種日常用品，發現每種都比去年價格上漲 5%，會否有人大聲疾呼：「生活成本增加了一倍！」

這就像路邊小販解釋他的兔子三明治如何能賣得這麼便宜。「我必須滲一些馬肉：一隻兔子的肉滲入一匹馬的肉。」

工會反對一位「聰明笨伯」老闆定義每小時平均工資：正常工時每小時$1.5，加班每小時 $2.25，周末加班每小時$3，共三小時得出平均每小時工資$2.25。這有意思嗎？

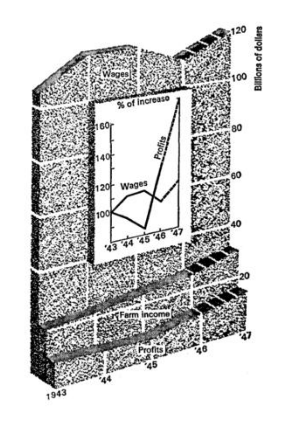
混淆「百分比 percentage」和「百分點 percentage point」是容易墮入的陷阱。如投資的利潤從去年的 3%攀升至今年的 6%，可以低調只是「增加三個百分點」，或是大事張揚「增加了百分之百」。特別是民意調查最常利用這種手法。



正態分佈的百分位數

百分位數 percentile 是統計術語，容易騙人。這基本上是將一組數據從小到大排序，並計算相應的累計百分位，某百分位所對應數據的值就稱為這百分位的百分位數。例如代數班有三百名學生，按各人成績排序，百分位數 99 是成績最佳前三名，其後三位是百分位數 98 ，依此類推。百分位數有奇怪而容易混淆的地方：百分位數 99 的三位學生的成績遠遠優秀於百分位數 90 的三位，而在百分位數 40 至 60 的幾十位學生成績可能幾乎相等。這是由於世事萬物的正態排序慣常呈鐘形曲線：最優最劣只佔少數，大多數趨向中位值。

偶爾統計人員發動內戰，旁觀者察覺到事有蹊蹺。美國鋼鐵工會為了爭取改善待遇，指出以 1939 年為基數，行業的生產力已大大提高，所以鋼鐵企業有能力加薪。工會沒有說明因為特別事故，1939 年的產量超低。企業的欺騙手法也不甘示弱，堅持員工的總薪資已有上升。這不是平均時薪，而是全體員工的總收入，其中包括許多早期以散工身份加入企業，後來轉為長工的人員；即使工資水平沒有上升，這麼多工人的收入必然會增加。



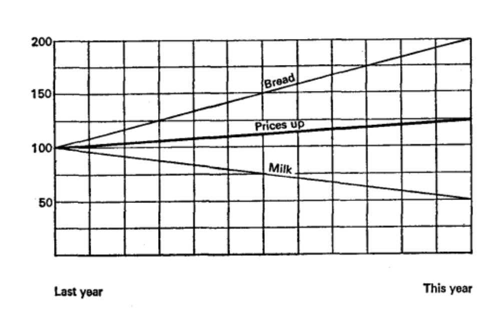
《時代》雜誌的圖形一向精益求精。這張插圖說明圖表可以是百寶袋，任由勞方資方隨意抽出所需的證據。這插圖其實是表達同樣數據的兩張插圖疊加一起。

方格圖顯示工資和利潤（以十億美元為格線比例），很明顯兩者都上升，而去年工資的增長是利潤的兩倍。以美元計，工資增長是利潤的六倍。巨大的通脹壓力似乎是來自工資。

白底插圖顯示工資和利潤增加的百分比。工資線相對平穩，利潤線大幅度向上。由此可見通脹壓力主要來自利潤。

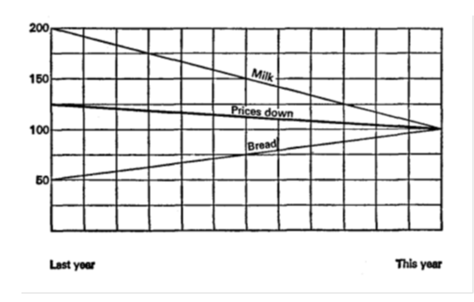
你可以得出自己的結論，或是更好的看到任何一方都不是通脹的主因。能夠及時簡單地指出爭論的主題不是表面的非黑即白，已經有助人們理解。

指數數字[52]至關重要，影響百萬受薪族的工資。因此要提醒各位這也是任人剪裁的。



以最簡單的例子為例：去年，牛奶每瓶 10 便士，麵包每個也是 10 便士。今年牛奶降價到 5 便士，麵包是 20 便士。這說明什麼？生活成本是漲了還是降了？還是沒有變化？

考慮以去年為基期，把當時價格作為 100%。由於牛奶價格減半(-50%)而麵包價格翻了一倍(+200%)；50 和 200 的平均值為 125，價格漲了 25%。



再試一次，以今年為基期。牛奶本來是現價的 200%，麵包是現價的 50%。去年價是今年的 125%。

為了證明成本水平沒有改變，簡單切換為幾何平均值[53]，並以兩個年份為基準。這少許有別於算術平均值，但也是完全合法，並在某些情況下是最有用和啟發。要得到三個數字的幾何平均值：各數相乘，得出立方根。四個數字取第四根，兩個數字取平方根。就是這樣。

以去年為基準，價格水平為 100。實際是每項乘以 100%，取其平方根，得出 100。以今年為基準，牛奶是去年價格 50%，麵包是 200%，200 乘以 50 得出 10,000；其平方根 100 即是幾何平均值。各項價格沒有上漲或下跌。

事實是儘管統計有數學基礎，但既是藝術，也是科學。在這範圍內有許多操作，甚至扭曲。通常情況下，統計學家必須選擇表達事實的方法，這是主觀的過程。在商業現實中，他不太可能選擇對己不利的方法，一如廣告撰稿人不會描繪贊助商的產品不堅實和不夠檔次，他會說輕巧和經濟。

即使是學術界可能也有偏差（可能無意識），特別想證明某這一點。

[52] Index number
[53] geometric average

這表明我們要三思統計材料，在報紙和書籍，雜誌和廣告的事實和數據。但隨意拒絕統計方法也是沒有意義。這就像拒絕閱讀，因為作家有時用文字來掩飾事實和關係，而不是披露公開。

補充材料

數據集的誤區

大量的數據才能得出有效的平均值，並準確預測趨勢。一萬人的數據優於一百人。只有 3-5 個數值的數據集，得出的結果並不真實。

數據集不僅要很大規模，也要很廣泛。地質學家調查沙漠數據，在沙漠十個不同地點收集 100 個數據，要比在同一地點收集 1,000 個數據更準確。

有兩個人，有一位雙腿截斷了。無論選擇哪一種平均值，只要不被看出只有兩個樣本，那麼就無法辯駁「人平均有一隻腳」的結論。

有些調查故意這樣做。例如，人口統計想要找出男性更傾向某種職業，那麼只需要調查男性人群。

一些小項調查經常錯誤地把控制集的調查結果等同普遍結果劃等號。小項調查沒有辦法調查廣泛、隨機的城市人口，學院調查經常方便地面向大學生人群，尤其是心理學測試實驗。即使調查報告說明情況，但新聞機構為了發表聳人聽聞的報道，往往把細節模糊，利用院校層次的調查結果來以偏概全。

使用不平衡的數據集撒謊的做法非常狡猾。技巧是把那些其實並不能相提並論的數據放在一起比較。例如，十萬人口的新城鎮在十年新增一萬人，比較原本只有十個居民的小村落在十年增多十個人，那麼就可以理直氣壯地總結小村落人口增長更快。

有時市場調查會利用這技巧來發表銷售數據。調查蘋果和橘子的銷量，但是調查到了一半橘子由於存貨不足賣光了，但調查依然繼續，那麼蘋果銷量就會遠遠高於橘子，即使蘋果並不是真的更受歡迎。

解讀調查數據的誤區

許多事物的因果關係涉及多個甚至無數的因素，調查往往不能孤立少數因素以設

計對照組研究。

另一方面，這些複雜關係又方便了調查從中撮出一些有利本身觀點的結論。常見的統計陷阱是調查測試包含大量應變項(dependent variable)，方便找出一個有利自己的似是而非的因果關係。

第十章　如何反駁統計的謊言

最後一章解釋如何看透虛假的統計，如何從中找出可信可用的統計。

不是所有眼見的統計訊息可以訴諸化學分析或踏實研究的誠實測試。以下五個簡單問題有助找出答案，避免受騙。

（一）誰的統計？

要尋找的第一個答案是偏見：進行調查和發表結果的一方有什麼動機？實驗室是為了理論，名聲還是收費而去證實什麼？報章是否追求銷路？勞資雙方是否要鼓吹某個工資水平？

留意故意的偏誤。這可能是直接的錯誤陳述，可能是模稜兩可的不明確聲明，可能是選擇有利數據和忽略不利數據，轉換測量單位（例如選擇有利的數據作比較），可能選用不適合的計量單位（例如採用平均數，而中位數能披露較翔實或更多訊息），以沒有說明的平均數掛羊頭賣狗肉。

公司宣布 3,003 人持有公司股票，平均持有 660 股。這是真實的數據，但沒有說明三位大股東已持有總股票數量四份之三，另外三千人共持有餘下的四份之一。

要留意無意的偏見，這往往是更危險。在 1928 年，許多統計學家和經濟學家發佈圖表和預測，證明經濟繁榮，無視經濟結構中的裂紋。

面對這些「證據」，至少要一看再看是誰發表這些統計數據，無論是聲名顯赫的政界、科學實驗室、甚至大學。報導引述：「某某大學研究發現…」，要注意的不是「某某大學研究發現…」，而是誰在引述，因為引述的結論往往是作者之言，不一定是某某大學的結論。

《芝加哥商業期刊》大事公告該期刊調查 169 間企業有關對抬高價格和囤積居奇的結果：三分之二企業宣布他們面對遠東地區的加價，是一如既往由企業吸收消化部份。期刊說（每遇上這些說話，要加倍留神！）：「調查顯示這些美國企業沒有追隨他人提價。」這是明顯的要質疑：「是誰這麼說？」由於期刊可被視為有利害關係，這也順延到第二個測試問題：

（二）他怎麼知道？

取樣

期刊相當取巧：事實是調查對象為 1,200 間公司，其中 9%回答沒有提價，5%有升價，86%沒有回答問卷。調查結果是基於有回答問卷的 14%。

要注意樣本偏差的證據，選錯樣本可能是無心，可能是有意。上文已提醒：樣本是否足夠的大，足以產生任何可靠的結論。

要同樣小心處理報導的相關性：相關性是否夠大，有重要的意義？是否有足夠的案例賦予任何意義？一般讀者不懂應用**顯著性檢驗**[54]來確定樣本是否足夠。但許多報導一眼就能看出（可能要花點時間）是否有足夠案例足以說服任何有理性的讀者。

（三）什麼不見了？

即使資訊來源響噹噹，如沒有明告有多少個案，已足以引起合理懷疑。同樣的，如提到關連性但沒有給出可靠性的計量（可能誤差，標準誤差），也足以引起合理懷疑。

提防平均值以及沒有指明的各種平均值，要知道在很多情況，平均數和中位數會有很大差別。

很多數字沒有意義，因為沒有比較。例如「蒙古症研究發現 2800 個案例超過一半的母親是 35 歲或以上」。除非知道婦女一般生兒育女的年齡，這說法沒有特別意義。很少人知道婦女一般生兒育女的年齡。

另一例子：「衛生部最近公佈的數據顯示在過去霧霾天氣的一周，死亡人數增加二百八十人。」死亡人數增加是否與霧霾有關？一般的死亡人數是多少？下一週的死亡人數會否減少？是否因為霧霾加速了某些人死亡？「死亡人數增加二百八十人」引人注意，但由於沒有其他數字比較，意義不大。

如只給出百分比而沒有原始數據，小心小心。很久之前，美國約翰霍金斯大學有一段有趣的報導：女大學生有 1/3 與教員共諧連理。驚人的百分比。原始數據說得清楚：許多年前，美國大學生只有極少數是女生；當年有三位女生，其中一人

[54] tests of significance

嫁給教師。

多年前，波士頓總商會的「優秀女性成就獎」宣稱：十六位名列名人錄的女士共有六十個學位和十八名子女。這些個人資料看來頗為紮實，但原本其中有兩位奇人，她們共有三十個學位，而其中一位有子女十二人。

留意指數有許多疏漏：可能是基數。勞工組織指出在經濟衰退後利潤和生產指數上升快於工資指數。指數沒錯，但沒有說明前者的基數較低，所以經濟復蘇時增加的百分比幾乎必然是較高。

有時指數的缺失是沒有說明導致變化的因素，有意或無意暗示是因為一些其他因素。今年二月的零售數字低於去年，但沒有指出去年的春節是在二月，今年在一月。

過去幾十年有關癌症死因的報告是誤導的，因為有許多外在因素：以前對癌症所知不多，死因往往列為「死因不明」；現在有更多死因解剖，診斷更可靠，醫療統計數據較齊全；現代人更長壽，更多人活到容易患上癌症的年齡。如果只看總死亡人數而不是死亡率，不要忽視現在的人口比以前更多的事實。

（四）是否有改變主題？

留意原始數據和結論之間是否被轉移，聲東擊西。

正如上文指出，更多呈報病例並不總是更多人染病。測驗民意的投票並不一定反映正式投票的結果。雜誌讀者的興趣調查不擔保他們會從頭到尾細讀文章。

某年，美國加州中央谷地呈報腦炎病例大幅度增加，是去年的三倍。很多居民感到震驚，把子女暫送外地。但死亡數字沒有很大改變；原來是州政府和聯邦政府開始投入資源解決這個長期問題；因為他們的努力，發現許多以往被忽略的低程度病例。

大家可能留意到在某段時間，報章特多報導某類型的罪案或事件，感覺是無日無之，但過不了多久又沉寂下來。如仔細追尋，相關的官方數字沒有增加。這只不過是有一兩位記者當其時特別多這方面的報導，其他記者不得不追隨其後。

英國公共工程部調查六千戶有代表性的家庭，發表報告：「英國男士在夏天平均每週沐浴 2 次，冬天 1.7 次；女性是 2 次和 1.5 次」；引來報章頭條報導英國男士每週沐浴次數多於女士。

這些數字要更令人信服，定要說明是平均數或中位數。然而，更嚴重的弱點是問題的主旨已經改變。調查真正發現的是「人們隨口回答他們的洗澡次數，而這不是反映現實」。這是相當隱私的問題，受訪者要顧全自己的面子（經常沐浴是良好的個人衛生習慣），對調查員給出的答案往往不是實際情況。

「離題」還有更多的品種變化。

《振興農業》調查發現美國農場比五年前增加了五十萬。這兩個相應的數字其實不是計量同樣的事情，因為調查局改變了農場的定義，新數據包括了舊定義不涵蓋的三十萬個農場。

人口普查發現奇怪的數據：例如三十五歲的人口不正常地多於三十四歲和三十六歲的人口。查究之下，發現數據是根據家人自報，他們傾向把歲數順便調整為方便的五的倍數。要解決這問題的方法是要求呈報準確的出生日期。

中國某大區「人口」是 28 萬，五年後升至 105 萬。這幅度的增長當然有問題，深究之下原來兩次調查是為了不同目的：第一次是稅務普查，第二個為了飢荒救濟。

美國也有一例。十年一度的人口普查發現 65 至 70 歲年齡組高於十年前的 55 至 60 年齡組。移民數字不能解釋這差異。主要原因是頗大數量的受訪者為了領取社會保障金而虛報年齡，也有可能是之前為了虛榮心而少報年齡。

美國參議員指責囚犯的住宿費用比市中心酒店還要昂貴，其實是混淆了囚犯的整體管理費用，這包括了監獄人員的薪俸。

各種事後孔明的廢話是暗地改變主題的另一方式。

還有許多「我是第一」的形式。幾乎任何事物都可以宣稱自己是第一，只要不是太特別的什麼。

當你考慮直接購買或分期付款，比較借錢成本會因為「改變主題」而難以比較。百分之六聽起來像百分之六，但可能不是真的如此。向銀行借貸 100 元，利率 6%，一年內每月清還利息約 3 元。但大多數汽車貸款標榜的「每百元利息六元」其利率實為雙倍，不容易明白。

更糟糕的是美國的冷凍食品計劃。粗心的買家被告知「6-10%」的數字。這聽起

來是利息，事實並非如此。這是還款的數字，更糟糕的是這往往是以六個月計算，不是一年。100 元價格的食品，每月還款 12 元，等同真正利率 48%。難怪有這麼多客戶拖欠，食品計劃要結束。

有時候會以語義來改變主題。《商業周刊》的報導：會計師決定「過剩」是討厭的詞語，提出企業資產負債表不再採用，改為「留存收益」或「固定資產增值」。

（五）是否有意義？

「是否有意義？」往往能夠把基於未經證實假設的整個繁瑣統計回歸應有地位。Rudolf Flesch 提出文章可讀性公式：簡單和客觀計算單詞和句子的長度。以數字取代無法估量的論述，以算術取代判斷，這是有吸引力的想法。至少僱用作家的人，如報紙出版商，甚至許多作家本身都應該注意。公式假設字詞的長度決定可讀性。這是否故意刁難，還有待證明。Robert A. Dufour 利用這公式評審一些文獻，頗為得心應手，有助判斷一篇文章、一本著作是否比較難讀。

許多統計數字表面上已是虛假，只因為數字的魔力令人忘卻了常識而蒙混過關。Leonard Engel 的多篇雜誌文章列舉了幾個醫療案例。

一個例子是著名的泌尿科專家計算美國有八千萬前列腺癌病例 - 足以涵蓋易感年齡組的每位男性！另一例是神經科醫生估計每十二名美國人有一人患有偏頭痛；因為偏頭痛佔慢性頭痛病例三分之一，這意味人人每一季度會患上失能性頭痛。還有一個例子是經常提到的二十萬宗多發性硬化症病，但死亡數據表明這種病例不會超過三至四萬宗。

關於修改社會保障法一直飽受各種形式的聲明；如未經仔細考證，這些聲明各有各的道理。論點之一是既然預期壽命只有約 63 年，退休年齡訂為 65 歲是虛假和欺詐行為，因為幾乎每個人都在這之前死亡。

只要看看你認識的人就可以反駁這論點。基本謬誤是這數字是指出生時的預期壽命，因此大約有一半嬰兒可以預期活到 65 歲。順便說一句，這數字來自 1939-41 年期間，已經過時但仍然使用。經過一代人後計算，目前的預測數字是 69.7 歲；這個新數字同樣愚蠢，幾乎每個人現在活到 65 歲。

多年前，一間大型家電公司的產品規劃是基於出生率下降，長久以來已被認為是理所當然。規劃要求重視小電器，適合公寓大小的冰箱。策劃者之一突然回歸常識：他放下圖形和圖表，轉而留意自己和同事、朋友、鄰居和舊同學，除了少數例外都有三、四個孩子或是計劃大家庭。這重新啟動沒有成見的調查和製圖 - 該

公司很快轉向最有利可圖的大戶型。

赫然精確的數字往往違背人們的常識。紐約市報紙報導一項研究：與家人同住的在職婦女每週生活所需是 40.13 元。任何有常識的讀者會意識到生活成本無法計算到最後一分錢。但是 40.13 元比「約 40 元」更動聽，更是可怕的誘惑。

外推法[55]是有用的，特別是所謂預測趨勢的占卜形式。看著這些數字和從中衍生的圖表，必須記住：至今的趨勢可能是事實，但未來趨勢只不過是有些見識的猜測而矣。隱含的意思是「一切因素不變」和「目前的趨勢繼續」，但世事偏偏不會保持不變，否則人生會很無聊。

不受控外推法的廢話，電視趨勢是例子。在最初五年，美國家庭的電視機數量以百倍增加。依此趨勢推論，再過五年會有幾千萬部，大概每家有四十部。

> 1948 年美國總統選戰預測是統計史的大笑話。選舉前的各項民意調查大多預測共和黨候選人 Tom Dewey 獲勝。結果是民主黨杜魯門得票 49%勝出。蓋洛普選舉預測被稱為「人類歷史上最公開的統計誤差」。
>
> 專家分析民調出現偏差的原因，結論有三：調查抽樣偏離了代表性、民調提早一星期結束，沒能反映最後時刻的民意變化，以及政治偏見妨害了編輯的客觀立場。當年報社老板多為共和黨人，報紙挺共和黨的當然較多。[56]

相對於一些未來人口預測，這已是準確的典範。近至 1938，總統的專家委員會深信美國人口永遠不會達到 1.4 億；十二年後這數字已是 1.52 億。這些可怕的低估源於假設趨勢將繼續沒有變化。

1874 年，馬克·吐溫總結了外推法的廢話：

> 在一百七十六年間，密西西比河下游縮短了 242 英里，即是每年平均縮短 1 1/3 英里。依此推論，一百萬年前的密西西比河下游足足有一百萬英里長，像釣魚桿伸出了墨西哥灣，也可以推論七百四十二年後，密西西比河下游將只有 1 3/4 英里。科學真有趣。只需投入少許事實就可以得出這樣的回報。

[55] Extrapolation
[56] 改寫自 http://hk.crntt.com/crn-webapp/mag/docDetail.jsp?coluid=36&docid=102284142&page=4

附錄　　　　　　香港大學民意調查的爭論

（自學書院註：在翻譯這本小書期間，香港正好有一場有關民意調查的筆戰，也正好印證民調和統計的重要意義和容易陷阱〔正反雙方皆如是〕。事緣香港特首[57]不是全民選舉產生，無從得知究竟有多少選民屬意他領導香港，於是定期民意調查是各方關注的寒暑表。香港大學民意研究計劃和香港中文大學亞太研究所的定期民調最為各方關注。現任香港特首梁振英自 2012 年 7 月就任以來，民望一直在所謂合格線(50)徘徊。為此，行政會議[58]議員張志剛向香港大學民意研究計劃發炮，引來一場不大不少的筆戰。奇怪的是亞太研究所的民調結論也是差不多的「不合格」，但梁粉〔梁振英粉絲〕沒有為此著墨。輯錄這幾篇文章頗多香港文體用語，請享用。）

港大民研發放特首及問責司局長民望數字

2014 年 3 月 11 日　〔香港大學民意研究計劃〕新聞公報

特別宣佈

在促進學術研究和理性討論的基礎上，香港大學民意研究計劃（民研計劃）今日在發放各項民望數字之餘，更加把關鍵原始數據上載到《香港大學民意網站》，包括特首評分、被訪者性別、年齡組別、以及加權指數。這種透明度，已經超過一般學術與專業要求，希望社會人士珍惜。學者專家使用及引用有關數據時，請按照學術慣例列明出處。

- 下載原始數據 ：2014 年 3 月 11 日公布之特首評分

公報簡要[59]

民研計劃在 2014 年 3 月 3 至 6 日期間，透過真實訪員以隨機抽樣方式，成功以電話訪問 1,017 名香港市民。調查顯示，特首梁振英的最新支持度評分為 47.5 分，支持率為 25%，反對率為 56%，民望淨值為負 31 個百分比，跟兩星期前變化不大。…根據民研計劃的標準，梁振英屬於「表現失敗」。 在 95%置信水平下，各項百分比的最高抽樣誤差為+/-4 個百分比，評分及支持率淨值誤差另計，調查的回應率為 66%。

注意事項：

[1]《香港大學民意網站》的網址為 http://hkupop.hku.hk，傳媒可到網站參閱調查細節。

[2] 調查之樣本為 1,017 個成功個案，並非 1,017 乘以回應率 65.9%，過去有不少傳媒在報導上犯了上述錯誤。

---

[57] 香港特別行政區行政長官（又稱特區首長、俗稱特首；英語：Chief Executive）

[58] Executive Council，即是特首「內閣」。

[59] 這項定期的民意調查涵蓋香港特區行政長官（特首）和主要官員的民望。為方便閱讀，附錄略去有關主要官員部份。

[3] 95%置信水平，是指倘若以不同隨機樣本重複進行有關調查 100 次，則 95 次的結果會在正負誤差之內。傳媒引用本調查的評分數字時，可以註明「在 95%置信水平下，各項評分誤差不超過+/-1.8，百分比誤差不超過+/-4%，淨值誤差不超過+/-6%」。 由於民研計劃在 2014 年引入「反覆多重加權法」處理數據，交接期間，各項數字變化的差異是否超過抽樣誤差，是基於同類加權方法處理後的結果計算。換言之，2014 年第一次所得數據是否與上次調查存在顯著差異，是基於兩組數據同樣經過反覆多重加權後作出的比較，而非單從公佈數字表面運算得來。

[4] 因為調查存在的抽樣誤差及處理數據的捨入過程，數字不能過份精確，合計數字亦未必完全準確。因此，傳媒在引用有關調查的百分比數字時，應避免使用小數點，在引用評分數字時，則可以使用一個小數點。

[5] 調查數據並非透過音頻互動系統取得，倘若調查機構以「電腦隨機抽樣電話訪問」或類似文字來掩飾音頻互動調查，是不專業的做法。

最新數據

民研計劃今日發放特首梁振英及各問責官員的最新民望數字。2014 年起，民研計劃把以往按照年齡及性別分佈進行的簡單加權方法，改良成為按照年齡、性別及教育程度（最高就讀程度）分佈的「反覆多重加權」方法調整數據。今天公佈的最新數據，是按照政府統計處提供之 2013 年底全港人口年齡及性別分佈初步統計數字，以及 2011 年人口普查收集之教育程度（最高就讀程度）分佈統計數字，以「反覆多重加權法」作出調整。現先列出最新調查的樣本資料：

調查日期	總樣本數	回應比率	最高百分比誤差[6]
3-6/3/2014	1,017	65.9%	+/-3%

[6] 有關誤差數字均以 95%置信水平及整體樣本計算。95%置信水平，是指倘若以不同隨機樣本重複進行有關調查 100 次，則 95 次的結果會在正負誤差之內。個別題目如果只涉及調查內若干次樣本，百分比誤差會相應增加。評分及支持率淨值誤差則會按照樣本評分及支持率淨值的分佈情況另行推算。

由於不同題目涉及調查內不同次樣本，誤差會相應變化。下列參考數表籠統列出樣本數目與最大抽樣誤差的關係，方便讀者掌握有關變化：

樣本數目（不論是總樣本或次樣本）	百分比誤差[7]（以最高值計）	樣本數目（不論是總樣本或次樣本）	百分比誤差[7]（以最高值計）
1,300	+/- 2.8 %	1,350	+/- 2.7 %
1,200	+/- 2.9 %	1,250	+/- 2.8 %
1,100	+/- 3.0 %	1,150	+/- 3.0 %
1,000	+/- 3.2 %	1,050	+/- 3.1 %

900	+/- 3.3 %	950	+/- 3.2 %
800	+/- 3.5 %	850	+/- 3.4 %
700	+/- 3.8 %	750	+/- 3.7 %
600	+/- 4.1 %	650	+/- 3.9 %
500	+/- 4.5 %	550	+/- 4.3 %
400	+/- 5.0 %	450	+/- 4.7 %

[7] 以 95%置信水平計。

以下是特首梁振英的最新民望數字：

調查日期	2-6/1/14	15/1/14[8]	18-22/1/14	4-6/2/14	17-20/2/14	3-6/3/14	最新變化
樣本基數	1,018	1,017	1,014	1,030	1,031	1,017	--
整體回應比率	66.5%	66.7%	67.6%	65.5%	67.8%	65.9%	--
最新結果	結果	結果	結果	結果	結果	結果及誤差[9]	--
特首梁振英評分	45.6	48.9[10]	47.0[10]	47.9	46.4	47.5+/-1.5	+1.1
梁振英出任特首支持率	27%	29%	29%	25%[10]	23%	25+/-3%	+2%
梁振英出任特首反對率	58%	53%[10]	54%	56%	56%	56+/-3%	--
支持率淨值	-31%	-24%[10]	-26%	-32%[10]	-33%	-31+/-5%	+2%

[8] 是次調查為施政報告即時調查，只問及特首評分及支持率。

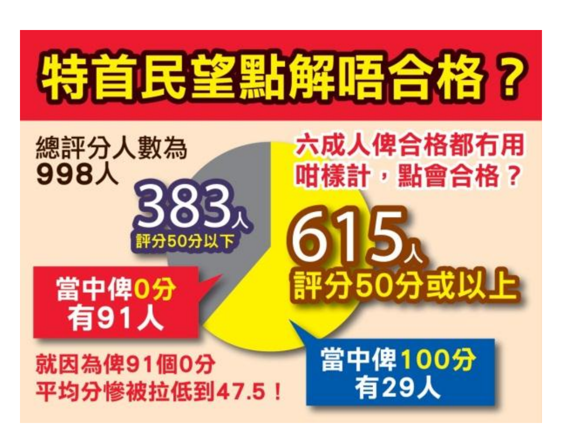
[9] 表中所有誤差數字以 95%置信水平計算。95%置信水平，即是指倘若以不同隨機樣本重複進行有關調查 100 次，則 95 次的結果會在正負誤差之內。傳媒引用上述數字時，可以註明「在 95%置信水平下，評分誤差不超過+/-1.5，百分比誤差不超過+/-3%，支持率淨值誤差不超過+/-5%」；以前調查的誤差數值請參閱網站。

[10] 該等變化在相同加權方法下超過在 95%置信水平的抽樣誤差，表示有關變化在統計學上表面成立。不過，數字變化在統計學上成立與否，並不等同有關變化的實際用途和意義。

【港人短評】解開特首民望「不合格」之謎

2014-03-14

港大民意研究計劃的民調早陣子引起連串質疑，未知是否有見及此，今次港大再度公布特首評分時，民意網站已出現所謂的「原始資料」，雖然相關檔案的格式要以特定軟件才能打開，但內裡所刊載的正正是評分分布數字。



民調應公正 做法須公平

依據港大最新的民調，以 100 分為滿分，特首僅獲 47.5 平均分，當然就被評為不合格了。然而，只要打開原始資料，就會發現 998 個評分者中，原來有多達 615 人、即逾 6 成人均給予特首 50 或以上的合格分數，其中更有 29 人給予 100 分；僅有 383 人給予 50 以下的評分。那麼，為何特首的評分又會不合格呢？最大的問題在於有 91 人個受訪者給予 0 分，就是這些極端評分，令特首的平均分大幅度拉低。

然而，這種意義甚為重要的評分分布，港大方面卻未有主動公布，而只是藏在民意網站的暗處，若非主動尋找及裝有特定軟件，根本無法知曉！這種藏頭露尾的安排，實在無法不令人懷疑民調背後的用意，即使不是存心誤導，但這又是否一個公正持平的民調機構所應採用的發布方式呢？

收集及公布數據 必須高度透明

要知道的是，民調機構如何採用、公布、以至運用收集回來的數字，對最終的民調結果又或市民觀感均起著決定性的影響。如此看來，香港確實有必要有更多獨立的機構進行民調，並要高度透明地公布收集到的數據，以助市民大眾通過比較獲得真象。

張志剛[60]：六成二給特首打 50 分或以上說明什麼！[61]

陳莊勤先生在 2 月 8 日於《明報》以〈沉默的螺旋〉為題撰文，對現時中大亞太所和港大民意研究計劃所做的特首評分提出質疑。重點就是機構只公布平均分，但打分分數的分佈卻不清楚，只靠一個平均分，根本無法知道事情的真象。而本人上周撰文，指出單靠一個平均數，其實就是瞎子摸象。一般的研究，除了平均數之外，多會公布眾數（最多人打的分數）、中位數，以及 50 分以上的比率。當時本人大膽推測，眾數和中位數都是 50，給特首打 50 分或以上的應該超過一半。文章見報當日，港大民意研究計劃也公布了最新的一次特首的評分，評分為 47.5，而港大也第一次以附錄形式把所有評分的原始數據同時公布，這也是解決了陳莊勤和本人過去一直提出的質疑。因為附錄必須要以 SPSS 軟件才能打開，一般媒體都不具備這種統計分析的專用軟件，所以沒有引起廣泛關注和報道。當我們打開這個原始數據檔案時，馬上真相大白。陳莊勤不用估，本人也不用猜。

港大首次公布所有原始數據

港大把給 0 分到 100 分的頻率全部公開，可以說是非常公開透明。為方便表述解釋，現把分數組合成 10 分一組，一共 10 組，評分分佈見附圖。

經運算之後，得出這樣的結果。平均分是 47.5，眾數是 50，中位數也是 50，給 50 分或以上的高達 61.8%。看完那些評分分佈以及這 4 個重要指標，我們不需要再瞎子摸象，象的形狀完全出現於我們眼前了！

平均分是 47.5，一般人的印象就是不合格！但如果看 50 分以上和以下的比例，在那 998 個給特首打了分數的人，有 28%的人打了 50 分，給 50 分以上的有 34%，那評 50 分以上的比率就是 62%，比 49 以及以下的 38%，多出一大截。當 62%香港市民給特首打 50 分或以上時，這是合格還是不合格？一些聳人聽聞的講法，例如民望破產之類，又從何說起。

把平均分拉到只有 47.5 分，最大的原因是大約有 9%的受訪者打了 0 分。本人之前撰文也解釋過，行政長官的施政有必然的兩面性，無論政策多好，都會有一些人不滿意。雙辣招有八成人支持，但還有兩成人反對，某程度是利益之爭，持有多個投資物業的人就不支持，地產經紀也不支持，迷信絕對自由市場的不支持。因為支持雙辣招而支持特首的，可能給 60 分，但反對雙辣招的就可能打 0 分。這種給行政首長的評分，就不能和讀書考試相比擬，資質良好、讀書用功的同學，

---

[60] 張志剛，香港行政會議（相等於內閣）成員，現任智庫組織「一國兩制研究中心」總裁。張志剛畢業於香港中文大學，分別獲授學士及碩士學位，文章常見於本港各大傳媒，著有《悲劇，悲香港》及《風雨聲中》等書。

[61] 原文刊載於《明報》2014 年 3 月 18 日

可以科科取得優異成績，甚至做 10A 狀元。但行政首長推行政策，一定有得有失，結果也只會把平均分拉向中間。如果不看分佈和其他指標，就只會以偏概全，甚至錯下判斷。

極端 10%主導輿情

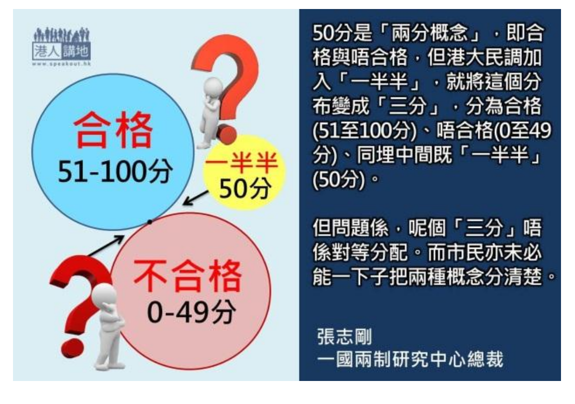
除了看那 50 分和以上佔了 62%的重要數據，我們不妨再把那 10 組的分數逐一研究，0 分到 9 分的有 10.5%，這是最極端反對梁先生的一群。但 10 到 19 分的卻只是 1.8%，20 到 29 分的也只有 3.9%。從分佈來看，這不算是正常的分佈，有點「惡之欲其死」的味道，到 30 和 40 分的兩組，才回復正常，逐步回升到 8.9%和 13.1%。

給 50 分或以上的分佈，就算是正常分佈最多的是 50 到 59 分，佔了 30.7%，愈高分數的比例愈低，逐步減少，沒有出現 10 分和 20 分組別近於斷層式的分佈。而這一成給予 0 到 9 分的群組，相信也是最主動發聲，最積極參與激烈行動的一群。當媒體的目光讓這一成人吸引着，所謂輿情，便傾向了這最極端的 10%。50 分以上的組群，他們相對平和理性，政府施政，他們心中有數，但沒有參與激進的意見表達活動，他們就成為了沉默的大多數。但當大學訪問員來電時，他們就把自己的評價說出，但不幸的是，他們的評分又給那 9%給零分的人拉低冲淡，如果沒有把所有得分公之於世的一日，這些沉默大多數的一群，永遠沒有見到「真象」的一日，也永遠讓那極端的 10%去主導輿情，和代表民情！

這種錯誤的代表，不僅是把民情扭曲，也形成了陳莊勤先生撰文中所提及的「白色恐怖的寒蟬效應」。支持梁先生的，支持特區政府的，都以為自己是少數，這令到他們變得沉默和冷漠，這也是反政府群體最希望見到的後果和現象。看完這堆港大公布的原始數字，真相大白於人前，支持梁先生的，支持特區政府的，不是少數！這說明過去一年半的政策走對頭，證明特區政府官員的「勤力用心」，市民是看在眼裡。

如果要正確的政策可以走下去，可以開花結果有成績，不僅是需要市民打一個分數，更是要他們表達意見，更是要他們站出來！

張志剛：50分應是「兩分概念」



對於港大民意研究計劃主任鍾庭耀解釋，民調中的50分代表「一半半」，即非合格，亦非不合格，一國兩制研究中心總裁張志剛表示，以0到100分給分本來是一個「兩分概念」，即合格與不合格，但港大民調加入了「一半半」，就將這個分布變成「三分」，即分為合格(51至100分)、不合格(0至49分)，以及中間的「一半半」(50分)。但他質疑，問題是，此「三分」並非「對等分配」，而市民亦未必能一下子把兩種概念分清楚。

練乙錚：打棍無效：網小子放倒「巨人」張志剛[62]

知識不等於力量，但如果缺乏知識，就可以很悲慘。無論在哪裡，若統治階級充斥不學無術之輩，社會大方向要出問題。這裡說的知識，當然不是「公婆皆可有理」的看法認知，而是客觀的學問。如果不僅是不學無術，還是別有心術的話，這個統治階級無可藥救。

臥虎藏龍

政改攤牌漸近，當權派集結力量圍攻鍾民調。先是政協委員、恒地副主席李家傑發飆，公開指摘鍾氏經常在關鍵時刻發布對特府或北京不利的民調結果，操弄民意，為反對派開路。跟着，梁派網站《港人講地》發表編輯室文章〈解開特首民望「不合格」之謎〉，指鍾氏在最近的一個關於特首民望的民調裡取巧運用數據說謊，把一個好端端成績亮麗的特首說成多數人視為「不合格」。然後，梁派悍將、行會成員張志剛高調發言並在本周二《明報》撰文，引用上述網文核心內容，質問鍾氏「六成二給特首打 50 分或以上說明什麼？」【註 1】

結果，「六成二給特首打 50 分或以上」說明了《港人講地》編輯室文章有「小小」搞錯了基本統計方法，而「國師」張志剛懵然不知（？）並加小手腳發揮，結果鬧大笑話。

最先指出《港人講地》文章和張志剛說法有好幾個嚴重初等錯誤的，是一篇又一篇的網上及新媒體文章，作者都懂統計，卻是傳統媒體裡不見經傳的業餘評論者，可謂小孩大衞打死巨人高利亞，亦可謂：網絡世界，臥虎藏龍。本文將這些材料整理，歸納所指出的謬誤，並加若干己見，給大家參考。

首先指出，張志剛文章（下稱「剛」文）的標題數字「62%」，與《港人講地》編輯室文章（下稱「講」文）同源，是一個發水或抽水幾近一倍的數字。「抽水」是指抽了民調回應者當中大批態度完全中立人士的水，把他們捆綁到梁特的支持者那邊，便成功創制出上述那個發水標題數字。過程中，還擅自替民調加上一個不適當的概念，對所導致的矛盾和足令梁特尷尬的結論卻諱莫如深。

張志剛的「62%」發水[63]近一倍

在港大鍾氏民調裡，特首「民望」數字的給定範圍是 0-100，內含 101 個整數，

---

[62] 《信報》2014 年 3 月 20 日

[63] 發水：滲水發大

50 分居中。訪問到的 998 個回應者當中，有 383 個給特首打的分數低於 50 分，280 個 50 分，335 個高於 50 分。鍾民調事先給受訪對象說明： 「0 分」為「絕對唔支持」,「50 分」定義是 「一半半」， 100 分則為「絕對支持」。

因此，對統計者而言，必須嚴格尊重那 280 個打 50 分者的中立態度，既不能把他們擺到 383 個不支持者那邊，亦不可將他們與 335 個梁特支持者放在一起；但是,「講」文捆綁抽水好自便，把打 50 分或以上的訪問對象加在一起（「一半半」+支持），一算：（280+335）/998 = 62%，好亮麗！

然後張志剛就用這個數字說事，雄辯地問：這個數字「是合格還是不合格？」這就有趣了。因為這個算法如果說明特首民望是「嚴重地合格」，那麼，我們同樣可以把那 280 個態度中立打 50 分的受訪者加到「不支持者」那邊（「一半半」+唔支持），算出（280+383）/998 = 66%。那不就表示梁特民望應該是「更嚴重地不合格」了麼？

矛盾兼尷尬！正如一篇網文題目所說：「你玩統計，統計玩你」。【註 2】
任何公平的統計人，不會像「講」文那樣，抽那些回應「一半半」的態度中立人士的水，而只會用 335/998 = 34%這個數字，代表在原始數據裡支持梁特的回應者比率。這個數字，固然比不上發水幾近一倍的「62%」，與不支持梁特的回應者比率 383/998 = 38%相比，也差一截。如此，張志剛更應該雄辯地問問自己：34%這個數字，「是合格還是不合格？」

為何說事者可如此便給，大抽態度中立人士的水？因為中間做了幾近無縫的概念轉移。

政治態度中立 → 「合格」→ 「支持」

大家如果留意，當可察覺「講」、「剛」二文其實歪曲了該次鍾民調裡的「50 分」的定義，把政治態度上的中立（「一半半」）巧妙地改成「合格」。然而這個民調裡的 50 分，並非是一個「合格線」。

「合格」的標準人人不同。例如，筆者當年念的大學，合格線因教授而異；念津貼小學的時候，學校的合格分數是 60%；中學則是 40%，入讀後，老父不滿名校的標準反而那麼低，筆者卻認為好得很，因為可減輕功課做不好給老父指罵時的「殺傷力」。

然而，更重要的是，合格和支持不支持，其實沒有必然關係—例如，某醫學院專科生以 40.1%的分數合格畢業，你支持不支持這位仁兄當你的心臟手術醫生？

「講」、「剛」二文先將「50 分」擅自定義為「合格」(與民調對象回答問卷時的指定意義不同),然後再把這個他們引入的「合格」概念等同民調裡的「支持」,這般偷換概念之後就可靜雞雞進行上述捆綁抽水。如此,「剛」文就可大剌剌地說:「評 50 分以上的比率就是 62%,比 49(分)以及以下的 38%,多出一大截。」(注意:「50 分以上的比率是 62%」起碼應該是「50 分或以上」罷?但連這個「或」字也省掉了。)如此逐步深入細緻地做群眾的思想擺布工作,不是第一次,大概也不會是最後一次。

事實上,港大民研計劃已再三聲明,「50 分」與「合格」完全無關,指的是態度上的中立。當然,可以有另外的民調專講合格不合格,但這個梁特民望民調本身不適宜講,硬要講,就會出現上面的既矛盾也讓梁特相當尷尬的結論。這個民調只研究特首民望的平均分數高低;得出一個平均分數之後,合格與否,讀者可憑個人喜好各自解讀。大概有些人,就算梁特民望拿個 1 分平均分,也會認為他是合格的;邏輯上,這沒有問題,但如果濫用民調原始資料特別炮製一個「62%」來說事,就有問題。

剔除給 0 分的!保留給 100 分的!

所說何事呢?原來,「講」、「剛」二文說,既有「62%」這個亮麗數字,而鍾民調最後竟把梁特的平均民望評分算為 47.5,必是因為鍾民調沒有把打 0 分的那些「極端分子」—即統計學上說的「離群數據」(outliers)—剔除。於是,他們就可結論:鍾民調不科學。這裡有三個問題。

首先,如果要剔除給 0 分者,也應該剔除給 100 分者罷?但張志剛口中振振有辭的那個發水「62%」,卻隱蔽地包含了 29 個「100 分」;這是「打茅波」。

其次,已經有專家算出,把回應分數最高和最低的 10%(含所有「0 分」和「100 分」)都剔除後,梁特民望平均值也好不了多少:48.1 分,救不了他;用張志剛的話說,依然「不合格」。如此,大動干戈為的顯然不是兩個平均分 48.1 - 47.5 = 0.6 分之差,因為「剛」文對此提都不提。那麼,要剔除 91 個「0 分極端分子」,目的何在?不外起哄,令不諳統計學的人「覺得」鍾民調無理。但請繼續看無理的是誰。

第三,回應分數值既限在整數 0 與 100 之間,而 0 與 100 分在民調裡都有清楚而具體定義,那麼,根本就不應剔除回應值為 0 或 100 的那些數據,因為那些數據已經不能算是「離群數據」,而是民調設計者特別指明、更要知道的數據;理論上,0 分甚至可能是對象回應中的一個「眾數」(mode)而意義尤其重要【註 3】。事實上,在該項民調裡,給 0 分的 91 個回應,佔 998 人的幾乎 10%,相當於給

50 分的 280 個回應人數的三分之一；這許多回應者，怎可以看成都是該從統計數字裡「槍斃」掉的呢？就看未加權的評分分布，我們也可以猜到，這個分布是雙眾數的（bimodal distribution），兩個眾數分別為 280 分和 0 分，因為的確有很多人對梁特極之不滿；若取消了這部分人的數據，那就不是今天的香港了。統計學不應、也不允許那樣搞出河蟹。

由此看出，不科學的不是鍾民調，而正正是《港人講地》編輯室和張志剛。
心術問題之外還有技術問題

「講」、「剛」二文，還犯了一個技術性錯誤：「62%」這個數字，是拿了鍾民調的原始數據做了小手腳就急不及待用來說事的結果，不知道人家有統計學的章法，就是對原始數據適當加權，之後才能用以作統計運算和分析。這裡說的「加權」指什麼？

大家知道，民調研究的對象人口總數太多，不能全部訪問，只能抽樣取板（sampling），但每一個隨機樣板中的個體特徵分布如年齡、性別等，都不能準確反映總人口中的已知分布，此即所謂的「樣板誤差」；如果所調查的民意（如對梁特的態度）與年齡、性別等特徵有關，樣板便需加工，而統計學用的標準加工工序，是一個加權工序。筆者借用近日一篇網上流傳很廣、署名 SweetSourPork（「咕嚕肉」）的《輔仁網》文章裡的具體解釋，稍作修改如下：

> 「如果今次電話訪問，有 41.5%嘅受訪者係男性，但係原來香港人口有 45.4%嘅人係男性，比受訪者入面嘅男性多，咁我哋就要將樣板入面嘅男性嘅比重加多啲，平衡番，等數據可以代表香港市民。」【註 4】

不做這個加權工序，樣板誤差可令民調的統計分析毫無意義。這是民調統計 ABC。「咕嚕肉」於是用了鍾民調的原始數據並作適當加權，重新再算一遍，證明鍾民調算出的梁特評分平均數 47.5 沒有錯，錯的是這裡又犯了基本統計方法大漏的《港人講地》和張志剛：那個已經包含抽水、概念僭建兼打茅波的「62%」，也是未經加權處理的（雖然因為前三個犯規動作太大太離譜，這第四個謬誤相對而言已顯得不那麼重要）。

大家看看，一個飽含四個大錯漏那麼豐富的「數字」，尊貴的行會成員視為至寶，雄辯滔滔用來攻擊對準鍾民調。那不是很可笑嗎？這種學養的人，放在本朝特府內外「智庫」裡打棍子很稱職，安插在行會，則說到底有損其他大部分成員的面子和心理。港大民意研究計劃成立於 1991 年，二十多年來，鍾民調的學術功架已經十分嫻熟，任憑當權派怎樣抹黑，亦不能把他撼倒。最近這次圍剿攻勢，網民當中的專家見招拆招，已經代為瓦解。正如筆者早前提到，鍾民調完全有資格

成為香港又一尊屹立不倒的圖騰。

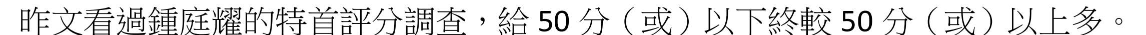
【註 1】李家傑言論見 http://zh.wikipedia.org/wiki/李家傑。《港人講地》編輯室文見 http://speakout.hk/index.php/2013-11-04-09-33-03/2013-12-21-08-43-26/1424-2014-03-14-10-38-16。張志剛文見 http://news.mingpao.com/20140318/msa.htm。

【註 2】見「陳電鋸」的文章 http://www.chainsawriot.com/archives/9292；此文用另一統計加權方法（iterative sample bootstrapping），算出梁特的平均評分為 46.3，比鍾民調的 47.5 稍低。

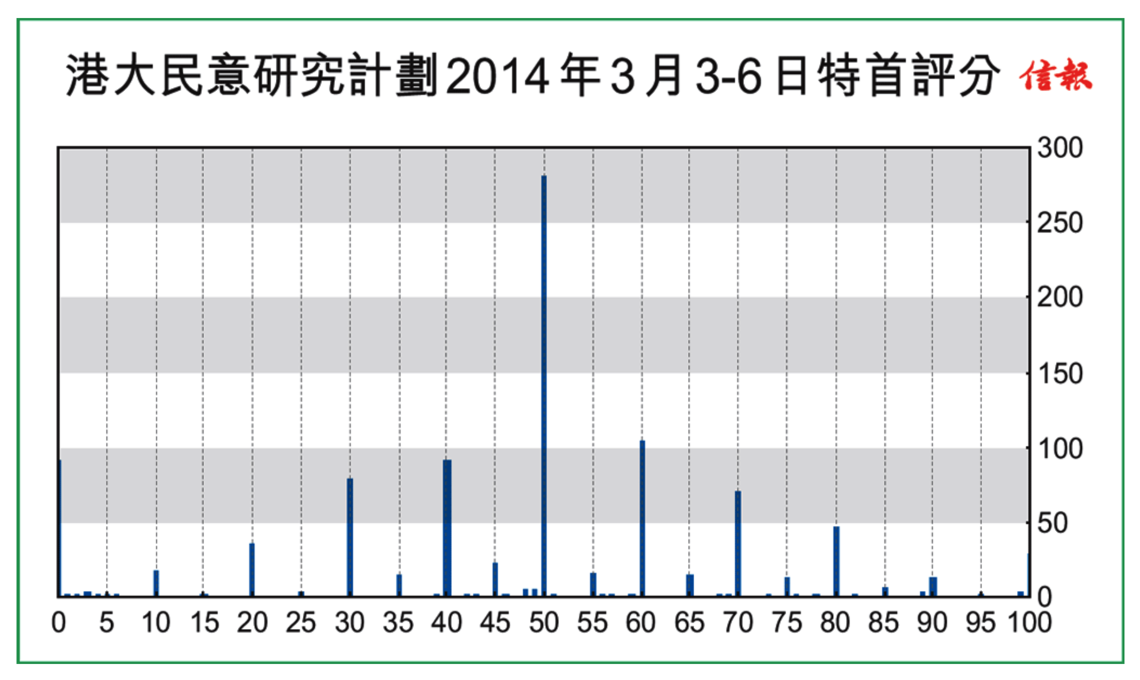
【註 3】關於離群數據，網文〈勿因蟲廢言〉 有很好的討論：http://aloneinthefart.blogspot.co.nz/2014/03/blog-post_15.html；作者指出，一般而言，問卷回應若不設有效頭尾限（例如 100 與 0）而是可以正負很大數以至無限的話，離群數據才有明顯的潛在不良作用，應該剔除。文章分析頭頭是道，明顯很在行；其上篇更值得看。

【註 4】「咕嚕肉」文章〈港大民研特首評分係「被拉高」還是「拉低」？〉，用典型香港話寫，解釋統計過程深入淺出，見 http://www.vjmedia.com.hk/articles/2014/03/15/66322。不過，文章的加權評分分布圖所表達的概念不對—應該是加權在人而不是加權在分，雖然算出的總平均分一樣是對的。

羅耕：低水準的批評[64]



昨文看過鍾庭耀的特首評分調查，給 50 分（或）以下終較 50 分（或）以上多。



說很多極端分子給 0 分嗎？一樣有不少給 100 分。難道全都要剔走嗎？觀乎分布，可能根本有些人想給超過 100 分，甚至有更多人想給負分，只是限於 0-100 無可奈何。如此 hit bound 的 tri-modal，用眾數（mode）表達是無甚意思的，因這很可能是（-∞,+∞）的正態分布。假使調查的 50 分水嶺改為 0 而兩端不限，大概未必會見到這三峰現象。

在平均（mean）、中位（median）及眾數三種中央趨勢描述而言，若是量化數據，最可取是平均。當平均有機會被極端數字大幅拉高／低時，才用中位，譬如入息分布。然而，特首評分限於 0-100，無極端數字，故不宜用中位。只有 interior multi-modal 下，以眾數表達多個中央趨勢才有意思。至於張志剛指的 inter-quartile range，更不必了。

數據是否正態分布，其實可以 Jarque-Bera normality test 測試，詳情可上維基看看。用原始數據不難算出，JB statistic 值達 386，顯然呈正態分布。

批評鍾庭耀的，看來要重新上基本統計課了。

---

[64] 《信報》2014 年 3 月 21 日

麥國華：民調科學與藝術[65]

回歸十多年，特首民望時常被傳媒打造成各具含義的大標題放在顯眼位置，製造話題。如果說傳媒為了吸引眼球而以文字渲染民調結果尚可理解的話，那麼一間理應中立的學術機構若真的選擇性公布某些調查數據，發布引導性結論，就實在令人為學術自由擔心。

近日，港大民意研究計劃遭揭發只公開有關特首支持度的「平均分」，而隱瞞原來有六成市民認為特首「及格」的事實，備受批評與質疑。然而，更讓人為之瞠目的是民研計劃負責人的反駁。他辯稱「從來不會用 50 分等於及格去解釋」，並稱 50 分只是代表「中間意見」。

支持程度本就是一種相當感官化的心理狀態，將其量化為具體數字，難免存在個人理解的因素。問卷設計者確可自行詮釋不同數字含義，此問卷亦將 50 分定義為「一半半」，然該負責人過往曾解釋「50 分以下等如不及格」，又何能自圓其說。加上某些自詡為香港良心的媒體也常以此為標準，疾呼特首民望不及格，大部分市民早被引導視 50 分為特首民望「及格」的界線。

面對質疑，該些媒體的反應更是令人心痛香港社會理智的流失。有媒體強調，揭出特首有 61%支持的是「梁粉」，暗示背後存在政治目的。一頂「梁粉」帽子就可否定一切證據事實。如此因人廢言，和文革時期不問觀點證據，單憑背景立場就批鬥廝殺有何不同？

很多平日鼓吹公義平等的「道德衞士」們，攻擊政府時高高舉起，現在面對涉嫌違反公義的事情卻又輕輕放低，彷彿事情只是橋下流水，其雙重標準也應予詬病。倘若被指民調欠缺公允的是中央政策組或建制派的民研機構，恐怕早已屍橫遍野了。只感嘆，民調可以選擇地公平，社會公義也可以選擇地分配。

捍衞學術自由

捍衞學術自由，政府、市民、政黨和學術界都有不可推卸的責任。民調的目的在於通過對大量樣本的問卷調查來客觀、精確地反映社會輿論或民意動向。民調結果會為政府所參考，從某種程度上可影響政府施政、市民心態及社會大環境。因此，市民有權利要求民研計劃本着嚴謹的學術研究態度進行調查，全面客觀地公布結果，讓公道回歸人心。遺憾的是，統計是一門科學，對統計數字的詮釋，卻是一門藝術。

[65] 《信報》2014 年 3 月 21 日

公說公有道，婆說婆有理？

「梁粉」批評如下：

依據港大最新的民調，以 100 分為滿分，特首僅獲 47.5 平均分，當然就被評為不合格了。然而，只要打開原始資料，就會發現 998 個評分者中，原來有多達 615 人、即逾 6 成人均給予特首 50 或以上的合格分數，其中更有 29 人給予 100 分；僅有 383 人給予 50 以下的評分。那麼，為何特首的評分又會不合格呢？最大的問題在於有 91 人個受訪者給予 0 分，就是這些極端評分，令特首的平均分大幅度拉低。

「主場新聞網站」及香港大學民意研究計劃研究經理李偉健則反駁：
評論指有 91 個 0 分樣本「拉低」平均分，沒有提到 29 個 100 分樣本同時會「拉高」平均分。港大民意計劃研究經理李偉健向《主場新聞》解釋，民望調查詢問受訪者給予官員 0 分至 100 分的評分，相信受訪者誠實回答，無論樣本是 0 分或是 100 分，都應納入計算，除非是 101 分，在數值範圍之外才會剔走。

李偉健強調，歷來民望調查同樣沿用這方法，公布按評分計算算術平均值（Arithmetic Mean），「沒有篩走特別低、特別高的評分。」

開門見山。我認為「梁粉」的批評有其道理，但其為己方所作辯解，一樣有問題。另一邊廂，「港大民研」的統計方法也有毛病。

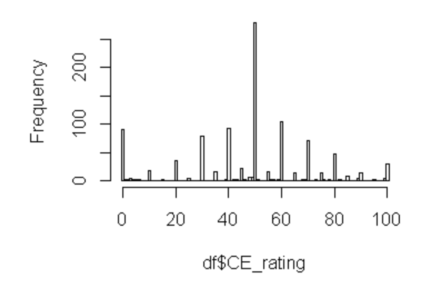
Lies, damned lies, and 梁粉's statistics

統計數字不會說謊，它有的只是統計偏差。說謊的，是運用它的人。"Lies, damned lies, and statistics" 這句名言，就是用來諷刺那些蓄意運用統計數字來製造假像的人。前述「梁粉」的批評，正好拿來作「統計語言偽術」的最佳範例。

從「特首民望調查」所得到的 998 個有效評分，平均分為 47.4（「港大民研」公布數字為 47.5，略有不同，這是因為他們按受訪者的統計特徵作加權平均），低於 50，但實際上 998 個分數當中，**有 615 個為 50 分或以上**……至此，梁粉都沒有說錯。然而，他們沒說的是：

998 個分數當中，**也有 663 個為 50 分或以下。**

感覺混淆嗎？或者這樣說吧，998 個分數當中，有 383 個低於 50 分，280 個等於 50 分，335 個高於 50 分。分數的分布如下：



從 0 到 100，共有一百零一個整數，而 50 正好居中。梁粉試圖以「50 分**或**以上」這個標準來描繪一個梁振英有超過六成人支持的景象，可是據他們的邏輯，我們同樣可以說，以「50 分**或**以下」這個標準來判斷的話，有超過六成人（而且這個「超過六成」的人數比起梁粉的「超過六成」更多）反對梁振英！

我不明白一眾梁粉何以如此介懷 47.5 這個只略低於 50 的數字。若是選舉的話，兩三個百分點也許是勝負關鍵，可是像印象分這種雖非玄學，卻也「不算精密科學」的東西，47.5 和 50，實在沒有分別。換了我是梁振英，看到如此數字，高興還來不及呢。

離群值與平均數

梁粉指出，998 個分數當中，有 91 個是 0 分，這些極端評分拉低了整體的平均數。這是正確的。「主場」卻反駁梁粉，說他們沒提及樣本當中亦有 29 個 100 分，會有拉高平均分的相反效果，也同樣正確，亦再一次顯示梁粉玩弄輸打贏要的統計語言偽術。

然而，撇除梁粉的拙劣技倆不談，若樣本中可能有不少「離群值」(outliers) 的話，到底我們應該如何估計統計母體群的平均數？

港大民研的李偉健指「無論樣本是 0 分或是 100 分，都應納入計算」。就一般統計調查來說，這是過時的做法（但此處有一個 catch，要押後談）。現代統計學認為「穩陣」(robust) 的做法，本網誌之前的書評其實已經提過，就是利用截尾平均 (trimmed mean)，也就是先截去最高和最低的 5-10% 數據，然後才計算平均數。

可是我們幾乎可以斷言，在「特首民望調查」中，無論用普通的算術平均，抑或用截尾平均，都不會有大分別。原因是一般來說，離群值最有殺傷力的情況，是母體群數字本身為「無界」(unbounded) 的時候。是項調查當中，有效的評分本身有界（只可介乎零至一百），離群值的影響通常不會太壞，故此梁粉的批評，抓不到統計學的重點。

實際上，若截去今次樣本當中，高低各一成的數據的話，得出來（未經加權）的截尾平均為 48.1，與樣本平均數 47.4 相去不遠。

尺度不同，分數如何換算？

這倒不是說「特首民望調查」無問題。印象中，港大民研所做的民意調查，大部份（例如立法會選舉的選前調查和 exit polls）都很紮實。然而此項「特首民望調查」，卻非常礙眼。我很想問鍾庭耀一句：How on earth is this rating meaningful?

單單叫受訪者為梁振英打個分數，已經很有問題。問卷只提過零分（「絕對唔支持」）、五十分（「一半半」）與一百分（「絕對支持」）的意義，中間的尺度 (scale)，人人卻不同細分。你我各給六十分，意思未必相同。你的分數如何換算成我的，完全木宰羊。現時港大民研的做法，實際上假設了所有人的評分尺度均一。由此引起的模型風險 (model risk)，無法評估。舉個例說，若你看到梁振英的「民望指數」比上月高，你可能以為他真的愈來愈受市民歡迎，但實情可能是他的民望無變，只是今個月的受訪者的評分尺度較寬鬆，對無甚特別感覺的官員，也傾向打一個高分而已。

就算是奧運體操項目，評分有較多稍為客觀的細項憑依（動作要求、難度、時限等等），仍不時惹人爭議，各人對特首表現的評分尺度，又怎可能大致一樣？

不知尺度，何論變化？

好了，就假設香港有一個平均的評分尺度吧。套用經濟語言來說，就當人人都用一個一致「市場評分尺度」好了，但為何我們可以計算平均分？平均數並不一定是有意義的。一半人給零分，另一半給一百分，借用時下流行語來說，是社會撕裂的狀況；所有人都打五十分，卻更似人人認命。兩種情況截然不同，平均分都是五十分，那麼五十分究竟是甚麼意思？

以上例子當然太極端，極端到與雷鼎鳴對堅尼系數的批評如出一轍。假若港大民研只是拿這個平均分來判斷粗略民情的話，上一段的批評是不適用的。問題是，港大民研對待這個平均數時，彷彿其精密數值或它幾個百分點的變化，有甚麼微言大義似的。然而，即使香港有一個「市場評分尺度」，我們仍不知道這個尺度是甚麼樣子。同樣是跌十分，從一百跌至九十分，是否跟六十跌至五十，或十跌至零同樣大鑊？木宰羊。五十分所代表的「一半半」，和「及格」是同樣意思嗎？木宰羊。不及格的話，甚麼分數才算民怨沸騰，很想梁振英辭職？木宰羊。

不知背後的評分尺度的話，再精密的數字都是沒用的。弄得好像很精密，反而令人誤以為該數字很科學，其細微變化很有意義。

離群值真是離群值嗎？
前面說過，以普通的算術平均來估計母體群平均數，乃過時做法。諷刺的是：

- 對「特首民望調查」來說，由於整把由零至一百分的量尺中，只有零、五十及一百有清晰意義，所以這三個分數，比其他分數可靠。
- 故此，**吊詭地，0 和 100 兩個離群值，反而不應剔除。**
- 結果梁粉針對離群值的批評，意外地不適用。
- 若硬要計算平均數，普通的算術平均，此處亦反而比截尾平均更恰當。

然而這不表示港大民研的做法正確。正正因為他們採用了語意不明的尺度，才造成這許多奇怪狀況。

結語一：less is more
如前述，港大民研的民意調查，一般都很紮實，但這項「特首民望調查」，用粵語來說的話，真係「畀位人插」。"Less is more" 這句說話聽來陳套，但此處適用。奉勸 Robert Chung，還是乾脆將問卷問題改成簡簡單單的「你想唔想梁振英繼續執政」之類好了，不要再搞那些懶細緻的評分吧。

結語二：廢話去死，自由萬歲
最後且談文字，不談統計。梁粉謂：

港大民意研究計劃的民調早陣子引起連串質疑，未知是否有見及此，今次港大再度公布特首評分時，民意網站已出現所謂的「原始資料」，**雖然相關檔案的格式要以特定軟件才能打開**，但內裡所刊載的正正是評分分布數字。
這不是廢話嗎？有甚麼檔案是任何軟件都可以打開的呢？何況所謂「特定軟件」和檔案格式，也不過是統計佬慣用的 SPSS 與它的 sav 格式吧。不想付鈔的朋友，可用免費的自由軟件 R 打開有關檔案。

相關網頁

- The R project for statistical computing
- 2014 年 3 月 11 日 新聞公報；香港大學民意研究計劃
- 下載原始數據（SPSS 的 sav 格式）：2014 年 3 月 11 日公布之特首評分
- 民情指數方法說明 (pdf)；香港大學民意研究計劃

伸延閱讀

- 電鋸，你玩統計，統計玩你：「問題根本不在於 0 和 100 等等 outliers ，而是佔人口比重較多的組群對梁振英評分較低。」

請鍾庭耀回應 請關焯照澄清／文︰ 張志剛

(明報) 2014 年 03 月 25 日

由前周鍾庭耀公布了特首評分的原始數據之後，就引起廣泛的分析和討論，這其實是好事。學術機構的行為，理應面對公眾批評，不要隨便就以「抹黑」和「打壓」視之。而關焯照先生等也寫了一篇專文，提出不同意見，個人在此嘗試把事情詳細再分析一遍。關先生和其他有興趣的人士可以詳細閱讀思考，往後可以再作交流或者當面討論。

整件事件似是複雜，但如作有條理的梳理，其實不難掌握。關鍵是鍾庭耀的特首評分，有沒有合格的概念和應用。此關鍵一解，往後就是大路一條。

鍾庭耀在 3 月 19 日接受《信報》訪問，指出「50 分是中位數，不能演繹成正向或負向數字，從來不能說 50 分合格」。

鍾庭耀的解釋，涉及兩個問題，一是這種評分，有沒有合格嬰ㄕ X 格的概念。二是如果有，又應該幾多分合格。

鍾庭耀的評分，其實做了很長歷史，太遠的不說，就從回歸談起，也有 17 年。這 17 年來，媒體從來都以合格嬰ㄕ X 格的概念來報道特首評分，而且都以 50 分為合格。香港媒體事業發達，每次數字一出，電視、電台、報章都踴躍報道，這合格嬰ㄕ X 格詞語，出現起碼 100 次。鍾庭耀每月起碼做一次調查，1 年 12 次，加起來就過千次。17 年來，少說也報了一兩萬次。如果鍾庭耀認為這個調查根本沒有合格嬰ㄕ X 格的概念，那在過去 1 萬多次的報道，鍾庭耀為什麼不挺身而出、撥亂反正？就在前周公布原始數據之後，得出評 50 分或以上有六成二人的結果，鍾庭耀才急忙表態，認為沒有合格不合格，又或者 50 分不能視為合格之說。

曾被引述 50 分為及格水平

香港的記者、編輯，多是有識之士，他們一個錯不奇，個個都出錯？他們視 50 分為合格，固然是憑自己的固有認知，而鍾庭耀自己也有不可推卸的責任。因為他仙人指路，他本人就是如此演繹。本人的一位同事用了一個下午的時間，在慧科電子剪報搜尋過去 10 多年的相關報道，找到以下這些材料。請記着，這些報道是直接經訪問引述或直述鍾庭耀的分析，而不是媒體自己的報道。如果只計媒體報道，那是成千上萬，不必在慧科電子剪報搜尋。

《蘋果日報》2010 年 8 月 11 日：「民意研究計劃總監鍾庭耀分析，按曾蔭權的民望表現而論，他的民望屬『表現失敗』。雖然他的評分有輕微上升，仍可以維持在略高於 50 分的及格水平。」

《頭條日報》2010 年 7 月 28 日：「該研究計劃總監鍾庭耀表示，雖然曾蔭權評分脫離肥佬行列。」（註：評分為 50.3 分）

《星島日報》2004 年 10 月 13 日：「鍾庭耀認為他（楊永強）的支持度保持穩定，比其歷史低位 39.4 分高出很多，但仍未達到 50 分的及格水平。」

《星島日報》2004 年 9 月 29 日：「鍾庭耀分析，調查結果顯示董建華的民望評分兩年來首次重上 50 分水平。」

《信報》2003 年 9 月 10 日：「鍾庭耀指出……孫明揚……楊永強……林瑞麟……馬時亨全數低於 50 分的及格水平。」

《明報》2003 年 8 月 13 日：「鍾庭耀分析：『……餘下 12 個問責官員中只有 4 個不及 50 分，算是初步走出管治危機。』」

《明報》2003 年 1 月 29 日：「鍾庭耀指出，特首評分自去年 8 月起已連續半年處於不及格水平……連續半年處於 50 分以下。」

另外慧科電子剪報顯示 2003 年 9 月 24 日和 2004 年 4 月 14 日的《蘋果日報》，在為特首和主要官員評分製表時，分別出現「註：評分以 50 分及格」（2003 年 9 月 24 日）、「註：評分由 0 至 100 分，50 分及格」（2004 年 4 月 14 日）等字樣，並且都寫明「資料來源：港大民意網站」。

鍾庭耀 1997 年 7 月出版的《民意快訊》第 11 期，在總結港督彭定康的支持度評分時表示：「整體而言，彭定康所得的分數一直能夠維持在 50 分的合格分數以上，反映彭定康在市民心目中的形象尚算不俗。」據港大民意網站介紹，無論是對回歸前的港督，還是回歸後的特首，支持度評分的提問方式是一樣的。

任何稍懂中文的人，也可以從上述的引述，清楚理解，這套評分方法是： 0 至 100 分，50 分為合格。講了千次萬次，鍾庭耀自己也是如是說。今日被翻出有六成二的人給了梁振英先生合格的分數，就走出來完全推翻過去 17 年的定義，作為香港大學的民意調查機構，鍾庭耀是不是要正式回應？

看完以上的引述，相信已經可以解答了關焯照先生的問題，但為求詳細交代，以

下再作進一步的分析。關先生等 3 人是懂得統計的人士，請 3 位首先思考並回答一個問題：鍾庭耀的評分，是歸類為定序（Ordinal）還是定距（Interval）的問題？所謂定序，通常是 3 項式選擇，回應者獨立挑選，只能每選項獨立計算頻率，選項之間也不存在空間可供選擇。中大在 2012 年初對候任行政長官支持度作調查，就提供了 3 個選項：不支持、普通嘢@半半、支持，這 3 個就是回應者可選的答案。在電腦運算時是用代碼，但運算後出來的答案結果仍然是不支持、普通嘢@半半、支持。如果是定序（Ordinal）的問題，我當然不能把一半半的歸類為支持，這是不能接受的錯誤，這種方法也同時不能相互運算，所以不會有平均分這結果。

看鍾庭耀問卷的問題，是典型的定距（Interval）的問題。0 至 100 是連續，不是獨立方塊。數字可以相互運算，所以有平均分的出現。如果關先生用 SPSS 查看鍾庭耀的原始數據，可以發現答案只是出現 0 至 100 分，從來沒有不支持、一半半、支持的字樣。這 3 組字不是答案，而只是用來向受訪者解釋 0 至 100 分的方向和意義。這個所謂一半半，在統計學上，和上述中大那個一半半，兩者完全不同意義。在定序（Ordinal）裏，一半半是獨立成章，本身就是答案。但在定距（Interval）中，50 分就是 50 分。而一般人對 50 分是合格分的印象已是根深柢固，早有定論。再加上媒體的報道，以及鍾庭耀自己也不斷解讀 50 分為合格分，所以本人以 50 分為合格分起點，向上計算得出 62%之數，又有何問題？如果真的要重回一半半的本來意義，那就只能用回中大那個問題，一半半獨立成章。但如果用 3 選項而不打分數，又無法製造「民望肥佬」的形象！

「平分春色」欠基礎

此外，關先生也提出把給 50 分的頻數一分為二，一半撥入支持，一半撥入反對，平分春色。

關先生這種做法，是完全混亂了取態上的一半半，和人數上的一半半。真的要知道給一半半的回應者的最後取態，就只能在訪問中再追問一條問題：「如果沒有一半半可選，那是會投入支持，還是投入反對？」另有一可能就是棄權不選。轉投的比例，根本無從得知，可能是八對二，也可能是三對七，我們憑什麼基礎去假設五成對五成？推論可以接受，但總要有一些基礎，例如參考其他兩分法民調的結果，而不可以隨意一分為二，這點希望關先生可以澄清。歸根究柢，我們必須清楚評分本身就有合格嘢ㄕ X 格的概念。而且一定有一個劃分點（cut-off point），而沒有中間形態。合格就合格，不合格就不合格，剛剛合格的下一個分數就是不合格，就是這麼簡單。

後記：默書拿 50 分的兒子問媽媽：「媽媽，我合格定唔合格？如果 50 分不算是合格，由 51 分才算，那 50 分又算什麼？又是合格，又是不合格？不能算是合格，

又不能算是不合格？」幾經折騰，媽媽最後無奈叫兒子：「你去問鍾 sir！」這時，妹妹跑過來告訴媽媽：「我默書也是 50 分，合格還是不合格？」媽媽喜形於色回答：「你哋一個合格，一個不合格。」（文章僅代表個人立場）

〈潮池 Blog〉**畫出腸民調之子矛子盾計**

不勝其煩，有關特首民望調查的爭論，無奈繼續。

港大民意研究計劃負責人鍾庭耀澄清，50 分在特首評分中，在問卷問題上，定義為「一半半」，統計學上屬「中間數」，不應視 50 分為「合格」或「不合格」(詳見〈畫出腸民調之一池渾水 〉)，張志剛在《明報》一文〈請鍾庭耀回應，請關焯照澄清〉，試圖以子之矛，攻子之盾，謂多年來，報章最少九次引述鍾庭耀形容「50 分為及格水平」，以證鍾庭耀自打嘴巴。

實情如何呢？

因為要準備是日香港電台《自由風自由 phone》節目，筆者用「慧科搜索」，複核了該文九個試圖指控鍾庭耀自打嘴巴的「例證」，功課已做，樂意公諸同好。

文字的確存在，不過……

(如果大家覺得好煩，請跳過以下二十三段，從尾六段開始看總結就可以了。)

(以下九「例證」引自張的文章)

「例證一」:《蘋果日報》2010 年 8 月 11 日：「民意研究計劃總監鍾庭耀分析，按曾蔭權 的民望表現而論，他的民望屬『表現失敗』。雖然他的評分有輕微上升，仍可以維持在略高於 50 分的及格水平。」

評：當天共有六份報章有引述鍾庭耀分析，只有《蘋果日報》提到他說「仍可以維持在略高於 50 分的及格水平」。1. 有可能是記者引述不精準，也有可能是鍾庭耀這樣說。2. 按前文後理，「仍可以維持在略高於 50 分的及格水平」有歧義，可詮釋為「50 分」是及格水平或「略高於 50 分」是及格水平。

「例證二」:《頭條日報》2010 年 7 月 28 日：「該研究計劃總監鍾庭耀表示，雖然曾蔭權評分脫離肥佬行列。」(註：評分為 50.3 分)

評：「脫離肥佬行列」，如何詮釋為「50 分為及格水平」？

「例證三」：《星島日報》2004 年 10 月 13 日：「鍾庭耀認為他(楊永強)的

支持度保持穩定，比其歷史低位 39.4 分高出很多，但仍未達到 50 分的及格水平。」

評：當天共有八份報章有引述鍾庭耀分析，只有《星島日報》引述鍾庭耀就樣說。有可能是記者引述不精準，也有可能是鍾確實這樣說過，難證實。中文大學的同類調查以五十分為「及格」，有可能令少部分記者也詮釋港大民研調查的五十分為「及格」。

「例證四」：《星島日報》2004 年 9 月 29 日：「鍾庭耀分析，調查結果顯示董建華的民望評分兩年來首次重上 50 分水平。」

評：當天共有十份報章有引述鍾庭耀分析，都有類似字眼，但「重上 50 分水平」，不可能解讀為「50 分為及格」的意思。正如評分「重上 60 分水平」，不可能解讀為「60 分為及格」。

「例證五」：《信報》2003 年 9 月 10 日：「鍾庭耀指出……孫明揚……楊永強……林瑞麟……馬時亨全數低於 50 分的及格水平。」

評：上段引述有很多省略號，原文是這樣的：

「鍾庭耀指出，市民對財政司司長唐英年及保安局局長李少光的評價相當不俗，可見人事更替似乎可以為政府帶來一點好處。不過，接替唐英年出任工商及科技局局長的曾俊華由於市民認知率不足三成而不獲排名。

房屋及規劃地政局局長孫明揚、衞生福利及食物局局長楊永強、政制事務局局長林瑞麟和財經事務局局長馬時亨全數低於五十分的及格水平，以林瑞麟及馬時亨最低分，分別有四十三分及四十二點九分。」

正常新聞寫法，很明顯最後一段並非引述鍾庭耀，「五十分的及格水平」屬記者自己的詮釋。「例證五」的省略號省得太多了。把兩段文字砌埋一齊，改變了意思，這就叫「斷章取義」。

「例證六」：《明報》2003 年 8 月 13 日：「鍾庭耀分析：『……餘下 12 個問責官員中只有 4 個不及 50 分，算是初步走出管治危機。』」

評：按當時詮釋的前文後理，鍾庭耀一直以 45 分為「信任危機線」，故有此說。而「不及 50 分」之講法，亦不能視「50 分為及格水平」。

「例證七」：《明報》2003 年 1 月 29 日：「鍾庭耀指出，特首評分自去年 8 月起

已連續半年處於不及格水平……連續半年處於 50 分以下。」

評：這是較離譜的一個引述，翻查原文，上述引文的省略號，省了三大段。原文第一段是「「港大民意網站」發現，特首董建華的民望，由 1 月中的 47.3 分跌至 1 月底的 45.2 分，下滑 2.1 分(若綜合其他數據，1 月平均分為 46.3 分，見圖)，再見歷史新低。民意研究計劃主任鍾庭耀指出，特首評分自去年 8 月起已連續半年處於不及格水平，反映政府有管治危機。」

然後隔了三段，才是「民意研究計劃主任鍾庭耀認為，特首民望自去年 8 月起，連續半年處於 50 分以下，並屢創新低，情況前所未有。」

而且，按鍾的說法，50 分以下，屬不及格水平 (50 分為一半半，50 分以上為及格)，此文與鍾的一貫講法無矛盾。如此拼湊證據，製造錯覺，唉。

「例證八」:另外慧科電子剪報顯示 2003 年 9 月 24 日和 2004 年 4 月 14 日的《蘋果日報》，在為特首和主要官員評分製表時，分別出現「註：評分以 50 分及格」（2003 年 9 月 24 日）、「註：評分由 0 至 100 分，50 分及格」（2004 年 4 月 14 日）等字樣，並且都寫明「資料來源：港大民意網站」。

評：不能排除「評分以 50 分及格」為記者的詮釋，在港大民意網站中，找不到「評分以 50 分及格」的字眼。找到的請告訴我。

「例證九」：鍾庭耀 1997 年 7 月出版的《民意快訊》第 11 期，在總結港督彭定康的支持度評分時表示：「整體而言，彭定康所得的分數一直能夠維持在 50 分的合格分數以上，反映彭定康在市民心目中的形象尚算不俗。」據港大民意網站介紹，無論是對回歸前的港督，還是回歸後的特首，支持度評分的提問方式是一樣的。

評：翻查港大民研出版的當期《民意快訊 》，確實清楚寫到 50 分為及格分數的說法。這是九個「例證」中，唯一一個清晰見到有「50 分為及格水平」的字眼。鍾庭耀如果要奉陪辯論下去的話，這點需要解釋。筆者意見，港大民研網站如大海一樣的歷史資料，只有一兩個矛盾位，「算係咁」。

長篇大論，真的唔好意思。總結：九個「例證」，五個為曲解、誤解或過分跳躍閱讀的錯解，三個有可能是記者自己的詮釋，只有一處 1997 年的說法出現矛盾。

張志剛與建制派的批評，一直針對港大民研計劃，其實中大也一直有同類型調查，為何不批判中大呢？他們要求要公開調查原始數據，港大民研自負盈虧，數

據屬學術資產，是日最新發展，港大民研發聲明，公開全部有關梁振英民望的原始數據，真的慷慨。其實，中央政策組也用公帑資助不少學者做研究，他們的研究成果，枉論公開原始數據，研究報告也只能於網上查閱到摘要。既有此「公開原始數據」的要求，是否公帑資助的研究，也應公開原始數據？

統計數據，應用 interval 還是 ordinal，各有優劣，50 分應如何定義與詮釋，本應屬於學術討論範疇，難分對錯，而且任何方式的詮釋，也只差兩三分，為何左報與建制輿論對一個學者頻密施襲了？大家何時對學術咁有興趣了？

事件風眼中的主角鍾庭耀，一直甚少正面回應各種批評，他最近在港台《傳媒透視》有一篇文章〈從國王的新衣的說起〉，詳細說了「國王的新衣」故事，文末有這樣兩段：

「國王沒有雅量，讒臣乘機取巧。先把小孩打成造反派，再把科學變歪理。然後口誅筆伐，肆意攻擊，製造白色恐怖，以為可以解決問題。誰知道，真理不會被改變。掩耳盜鈴，只會弄巧反拙。

面對來勢洶洶的攻擊，筆者並不急於回應。有助學術研究和公民社會發展的理性討論，筆者當然積極參與。對於那些不懷好意、借故詆毀的謾罵，就由它們在歷史洪流中消失好了。真理不在口舌之間，只要把事實紀錄下來，誰是誰非，歷史自有分曉。」

民調　真相此中尋　［關焯照、周文林、雷照盛］

蘋果日報　2014 年 3 月 26 日

港大民意研究計劃（下稱「港大民研」）的特首民調爭議越演越烈。網站「港人講地」和行會成員張志剛在這幾天仍在電子傳媒和報章發表批評，認為港大民研以評分計算民望的做法有問題。同時，將 50 分釐定為「一半半」可被一般市民視為合格分數，此外，將被訪者的評分劃分為「0 至 49 分」、「50」及「51 至 100 分」的概念，可能令問題含糊化。

首先，筆者寫這篇文章的目的是，（1）澄清一下做民調分析需要注意的地方，（2）希望避免民調結果的解讀產生誤解。

港人講地及張志剛猛烈批評的港大民研的民調問題是特首的支持度評分，其的內容是：「而家想請你用 0-100 分評價你對特首梁振英的支持度，0 分代表絕對不支持，100 分代表絕對支持，50 分代表一半半，你會畀幾多分梁振英呢？」

港大民研是採用統計學上常用的等距量表（Interval Scale）的方法去量度特首的支持度（由最低的 0 分至最高的 100 分）。這種做法的好處是從得分上了解到市民支持特首的「程度」（附圖）。大家可以細想，有兩位被訪者給予的分數是 51 分和 90 分，顯然，評 90 分的被訪者的支持度遠較評 51 分的被訪者為高，但如果採用港人講地和張志剛的提議方法去分組，以 50 分為中間點分界，然後將 0-49 分和 50-100 分別釐定為「不合格」和「合格」，讀者便不能看到這兩個評分的差異了。

港人講地和張志剛的做法是將 0 至 100 分的範圍變換為兩個不同組別，「合格」與「不合格」。如果用統計學的說法，他們是用一個順序量表（Ordinal Scale）去將數據分類——即是變為分類數據。如果用以上例子，51 分和 90 分是納入為同一組別（合格），但問題是 51 分和 90 分是分代表不同程度的支持，但在歸納組別過程（Aggregation）中，這種支持程度的差距便會被剔除，對研究者來說，這可視為流失了重要資料，最終令研究質量被拉低。

一個相關的難題是一旦採用港人講地和張志剛所提出的二元答案（合格和不合格）作為分析，在這情況下，問題的字眼和答案是需要修改。例如，問題可寫為：「你支不支持特首梁振英？」而答案分別是「支持」、「不支持」和「無意見」。一旦港大民研的問題重新改寫為港人講地和張志剛的問題格式，得出來結果（例如支持度的百分比）是極可能有差距，因為問題的本質和問法已不同，至於差距

在統計學上是否有明顯分別，這便要用適當的統計方法去驗證了。

最後，另一個爭論點是 50 分是否一個合格分。單以民調的問題措辭，筆者看不到港大民研有任何表示 50 分是一個合格分數。至於「一半半」，是一個中性詞彙，可解讀為「中間點」、「一般」、「普普通通」等。然而港人講地和張志剛堅持認為 50 分是一般人理解為合格分數，這只是他個人意見，正確與否，學界自有公論。

現在整個港大民研的民調爭議只是各說各話，猶如雞同鴨講。但筆者要指出，做學術研究是需要保持嚴謹態度，無論從民調內容、樣本的收集方法和統計分析均要達到起碼的學術水平，這才能令人信服。

關焯照　經濟學家、冠域商業及經濟研究中心主任
周文林　經濟學家、冠域商業及經濟研究中心高級研究員
雷照盛　統計學家、港大統計及精算學系講師、冠域商業及經濟研究中心研究員

盧先亞：特首的媽（一）

2014-3-28

前幾天看到了張志剛先生為了護主，在他報再次向鍾庭耀博士及挺身而出的關焯照博士，就民調一事「叫陣」，且在文中引述好些統計學的專業用語，例如甚麼等距 （interval）、有序（Ordinal）數據等等， 明顯就是要嚇唬外行人。我自問不學無術，未敢輕言反駁，所以特地請教我的一位學弟，現該說是一位學者。他與統計結緣廿多年，持有統計學博士學位，年少時甚至當過訪問員，及後任教統計課程，並主理多個大型統計調查及參與民調工作，現仍在這領域繼續研究，可知其醉心程度。

當我致電並道明來意，他努力嘗試透過電話解說，我越聽越唔知佢嗡乜，咁話晒都係學究嘛，當他亦然發覺話筒另端的「接收」 有問題，他說不如發個電郵以資說明，我自是求之不得。雖然我還得再三懇請他要寫得淺白入屋一些，而他亦同時叮囑我千祈「唔好開名」。我明白學院中人大都不愛拋頭露面，惟更清楚的是，若然無端拖他下水， 只怕鍾庭耀之外，又多一位統計專才遭受打壓，我又於心何忍。不過，跟手收到其洋洋數千字的鴻文更知， 其實佢根本就係想直斥痛罵張志剛！我又怎不玉成美事。惜原文太長，節錄之餘，還要分日刊出。以下是學弟的話，而括號內乃我後加：

張志剛先生，在此回應你在報刊所寫。特首也並不是我的兒子，我更不願作特首的媽！（誰又想天天捱罵呢！）一區之首亦不是小朋友默書考試！我不知道閣下對兒女要求如何，但大部分港媽亦不會接受仔女只拿 50 分，何況是特首要職！比方說，在職場上，誰會接受在工作上只有 50 分的下屬？怕早給炒掉了！（這點我可佐證）大部分有志氣有理想的人（與張先生無關），亦不會甘心跟隨能力只有 50 分的上司工作，沒前途的吧！所以請不要在 50 分上沾沾自喜，況且我們的特首在最新的港大民調中只得 47.5 分呢！

在張先生文中，論定港大民調問卷中的所謂支持程度是屬於等距(interval) 數據，原因是原始數據(raw data) 只記錄了 0 至 100 分，當中並沒有支持、一半半及不支持的字樣。這種論證確實粗疏！專業統計人員都知道，原始數據不能單獨使用，一定要參照編碼手冊(coding manual) 或問卷設計。舉例，問卷可能會包含一些有關出生地、職業、行業等問題，一般會用數字代碼記錄（例如 1 代表香港、2 內地及 3 其他地方），一來比較方便，亦同時大大減少電子檔案存量。如果不參照編碼手冊 （coding manual）或問卷設計，原始數據就出現不能解讀，甚或誤讀的情況。而張先生的論據只是簡單對號入座的誤讀罷了。

參考港大民調問卷，該問題是：「而家想請你用 0 至 100 分評價你對特首梁振英既支持程度，0 分代表絕對唔支持，100 分代表絕對支持，50 分代表一半半，你會俾幾多分特首梁振英呢？」自 90 年代起，港大民調一向是使用 CATI 系統（學弟列出全寫， 我從略），即是使用電腦抽選電話，自動撥號至接通，訪問員會準確依據電腦所示讀出問題再把受訪者答案輸入電腦，整個過程亦有主管在旁監聽以確保數據質素。所以每個受訪者亦會清楚明白 50 分代表一半半， 而不是代表合格，這是無可爭議的。

盧先亞：特首的媽（二）

2014-3-31

在討論甚麼是合格之前，首先要了解甚麼是支持程度。支持程度和考試測驗最大的分別是後者大多數有明確的評分標準，例如答對一題有 10 分，而合格標準則是老師或教授們的專業判斷。學術程度越高，合格標準就越嚴格，例如醫生、工程師等專業考試要求就很高，人命關天噢！所以考試分數大多是定義明確的集合（well-defined set） 。但在社會研究或行為科學等領域中，很多時要處理一些含糊不清、定義不明確的變數（variable） ，數學上稱為模糊集合（Fuzzy set ），例如快樂、情緒、生活滿足（life satisfaction）、工作動力 （work motivation）等等。一些社會學家、心理學家、計量心理學者 （psychometrician） 、教育學者就會以李克特量表（Likert Scale，下簡稱量表） 為這些模糊概念作簡單的量化描述，即是問卷常用的 5 級設計 ：

1. 非常同意
2. 同意
3. 既不是同意亦不是不同意（或作中立）
4. 不同意
5. 非常不同意

有些研究員會再把量表擴展為 7 級或更高級別，而港大民調只是把量表以 0 至 100 分表示，而 50 分則為 101 級量表的「一半半」！ 對照 5 級量表其實分別不大，只是支持及不支持兩方面被劃分得更仔細。值得注意的是，量表並非等距，即是 （4 不同意）並不是（2 同意）的兩倍，但一定對稱（symmetric）。同理，港大民調中所謂的支持程度，50 分亦不是 25 分的 2 倍，而用量表所計算出來的平均數亦只是一種中間趨勢的描述，這亦是對稱設計的結果。

那麼怎樣才叫合格？鍾博士講得很清楚， 在港大民調設計之中並沒有考慮這問

題！至於怎樣去訂立合格線，我建議可在港大民調中加入問題，例如問：你覺得作為一個特首，社會大眾對其支持程度（0 至 100 分） 應該（i）要達到幾多分以上才可以叫做合格（即 Pass）呢？（ii）要達到幾多分以上才可以叫做良（即 Pass with Credit）？ （iii）要達到幾多分以上才可以叫做優（即 Pass with Distinction）？ 另外， 亦可找來政治學及公共行政學的學者（經濟學者，尤其姓雷的，大可不必）們，為特首這職位定一些標準。當中並不一定只採用社會大眾的支持程度作唯一條件，同時可加入其他可測計量，例如 GDP 增長、堅尼系數、犯罪率、環保指標、新聞及言論自由指標等等。

我只想強調，特首是重要之職，合格並不足夠，香港作為一個現代化的國際城市，要有一個具傑出工作能力並獲大眾支持的特首方是王道。另外，張先生一再要鍾博士為過去傳媒的報道負責。這顯然不是統計問題，但我亦想請教張先生，有幾許公眾人物包括特首、司長、局長以致閣下又何曾會為傳媒的報導負責呢？梁振英 N 年前也說不會選特首，張先生曾幾何時亦公開讚揚港大民調中立專業。那張先生又如何對自己的言論負責？梁特首又是否要為自己反口食言負責呢？

事實上，民調是一項以統計學為基礎的社會研究專門科學，張先生可能並不是這方面的專才，那麼還請留待其他學者們討論交流。而張先生貴為行會成員，亦請不要重私忘公，免得引起社會大眾誤會行會打壓學術自由，那就相當不妙!

最後，我要向張先生表達敬意，你甘願接納與支持一個不足 50 分的特首，只因視特首如己出，把他當作兒子般看待，實有為人母親的偉大情操！（主席按：果然是溫良恭讓的學者，未句明明就是「他媽的」偉大！）

港大民調之統計學解讀《有涯小扎》

摘要：本文透過統計學分析方法，檢視近日輿論對港大民調中特首民望調查的批評及反駁，探討這些言論背後的統計學理據。本文作者認為，港大民調在抽樣方面十分嚴謹，但在設計問卷和演繹結果方面有值得商榷之處。本文又對港大民研所公布的原始數據進行了進一步分析，指出當中所蘊含的啟示，並據此提出建議。

引言

近日有關香港大學民意調查（下稱港大民調）的爭論甚囂塵上。港大民調是香港大學民意研究計劃（下稱港大民研）定期舉行的民調，由香港大學政治與公共行政學系的鍾庭耀主持。民調內容包括特首、政府、主要官員、議員民望，及其它社會指標等（《香港大學民意研究計劃》）。2014 年 2 月 8 日，民主黨黨員、律師陳莊勤在明報發表《沉默的螺旋》一文，批評港大民調以平均分來表達特首梁振英民望，結果易被極端數值影響，又以 50 分作為合格分數，並不全面。同時這些民調「本身並不單單在反映民意，也同時在以定期公布評分來塑造民意」（2 月 8 日明報陳莊勤《沉默的螺旋》）。3 月 4 日，在北京舉行的政協港澳聯組會議上，政協常委、恒基地產副主席李家傑點名批評鍾庭耀，指其主持的港大民調「總是在關鍵時候發表對中央政府、特區政府以至整個愛國愛港陣營十分不利的民意調查結果」，藉此「操弄民意」。他又認為鍾的民調不夠科學，卻是香港眾多民調機構中最具影響力的一個，必須盡快改變（3 月 5 日 AM730《李家傑批評鍾庭耀用民調為反對派造勢》）。鍾庭耀於同日發表書面聲明回應，指出其調查方法經得起學術考驗，「總會堅持科學透明的原則，從不遷就對方的政治背景或立場」，認為「如果把言論自由的憂慮，進一步擴大至學術自由的空間，是非常不智的做法。」他又歡迎任何人士討論民意研究工作，「只要是實事求是，客觀公正，便可集思廣益」（港大民研《關於政協委員李家傑於政協會議上有關「民意調查」的言論》）。

爭論焦點

陳、李二人的批評引起了廣泛關注。有論者從政治立場和動機立論（如 3 月 17 日文滙報文平理《「鍾氏民調」真的是學術嗎？》、3 月 18 日蘋果日報李怡《攻民調為扼殺民意》），本文對此無意涉獵。另有論者從統計學角度評論鍾的研究方法。行政會議成員張志剛在電台節目稱，鍾庭耀曾經多次提到 50 分是合格水平，認為他有需要向公眾交代（3 月 20 日商業電台《張志剛指鍾庭耀多次提及五十分屬合格》）。他又認為，在極端評分的影響下，用平均分來評核梁振英表現，猶如瞎子摸象，普通人亦難以理解 50 分是否合格水平。若 50 分屬於不合格，港大應清楚說明，並解釋何謂支持度評分合格或不合格 （3 月 21 日大公報《張

志剛促鍾庭耀交代 民望 50 分是否合格》）。陳莊勤則指出，「在一般人心目中，50 分這及格分具有非常重要的象徵意義」，但如果只公布平均分而不公布各評分的人數分布，便是不完整的民調結果公布。以今次民調為例，61.8%受訪者給予合格分數，38.2% 給予不合格分數，跟兩大民研／民調機構定期公布以平均分均多數低於 50 分所顯示的民情相去甚遠（2 月 8 日明報陳莊勤《沉默的螺旋》、3 月 20 日明報陳莊勤《再談民調》）。網站「港人講地」亦提出類似論點，指出整體平均分被 0 分的「極端評分」拉低，令梁振英支持度被低估，認為應取中位數更佳。過往多年的新聞報道都把 50 分演繹為及格分數，港大民研亦未有澄清，令市民累積了「50 分等同合格」的印象。又批評港大以 SPSS 格式發佈原始數據，必須裝有特定軟件才能開啟（3 月 14 日港人講地《解開特首民望「不合格」之謎》、3 月 20 日港人講地《有關港大民調的幾個疑問：覆練乙錚及關焯照兩位學者》）。公民黨黨員、港大法律學院院長陳文敏認為，剔除極端數據是普遍做法，因為更能反映現實（YouTube 視頻《公民黨港大法律學院院長陳文敏都覺得鍾庭耀的民調做法不是專業手法》）。中大亞太研究所研究員鄭宏泰稱，港大民調的 50 分沒有正面意思，不能視為合格，與中大民調講明 50 分及格並不相同。但 0 分亦是表達出某類民意，從政者應予注意（3 月 20 日明報《特首民望 50 分意義 中大「及格」 港大「一半半」》）。

因應批評，鍾庭耀在港大民研網站重貼了 2003 年的兩篇文章，解讀特首民望調查的設計（《「特首民望新解」、「問責官員如何向民意問責？」》）。文章指出，55 分的支持度大約等如假想投票中的 45%的「得票率」，50 分的支持度則可化成大約 30%的「得票率」，45 分大概會轉化為 20%，而 40 分大概會化成 10%至 15%左右。其後，鍾又在出席一個論壇時回應，指使用平均分是國際常用標準。而 50 分只是一個中性的分數，沒有所謂合格不合格。至於開啟 SPSS 格式檔案的軟件，在大學可以免費下載，他相信任何一個專業研究機構都有相關軟件（3 月 15 日商業電台《鍾庭耀指國際間最常使用平均分作研究結果》）。前中大經濟學教授、現職冠域商業及經濟研究中心的關焯照，聯同經濟學家周文林、統計學家雷照盛等撰文，指出根據問題的措辭，50 分只是代表「一半半」，沒有任何暗示這是一個合格的最低門檻。如果把 50 分歸入合格，會得出 61.8%的人給了合格分數。但如果把 50 分歸入不合格，會得出 66.4%的人給了不合格分數，兩者結果相反。解決方法是把一半評 50 分的人歸入 0-50 分一組，另一半歸入 50-100 分一組，結果是有 52.4%的人給了 0-50 分，反映特首的支持度評分不是太理想。他們同意一旦出現很多人選擇極高或極低評分，平均分不是最好的指標，建議同時公佈中位數和眾數，或剔除極高或低評分部份，計算「截尾均值」。但他們亦認為，極高和極低的評分也是重要的統計資料，不能忽略（3 月 20 日蘋果日報關焯照、周文林、雷照盛《民調小學雞》）。傳媒工作者練乙錚則指，港大民調的特首民望評分由 0 至 100，即有 101 個整數，50 分居中，故此應尊重給予 50 分者的中立態度，而非把 50 分理解為支持梁振英。至於 0 分與 100 分，在港大

民調中都有清楚而具體的定義，不應剔除。若真要剔除 0 分，亦應同時剔除 100 分。即使剔除了，平均值仍是低於 50 分（3 月 20 日信報練乙錚《打棍無效：網小子放倒「巨人」張志剛》）。

下表總結了兩方面的言論：

	批評	反駁
平均分與極端評分	▪ 整體平均分被極端評分拉低，低估特首支持度。(陳莊勤、港人講地) ▪ 剔除極端數據是普遍做法，更能反映現實。(陳文敏) ▪ 一旦出現很多人選擇極高或極低評分，平均分不是最好的指標。可同時公佈中位數和眾數，或剔除極高或低評分部份，計算「截尾均值」。(關焯照等) ▪ 類似 0 分或 100 分的的極端評分將會愈來愈多，因此不能單單公佈平均分，可以中位數代之。(港人講地)	▪ 使用平均分是國際常用標準。(鍾庭耀) ▪ 0 分亦表達出某類民意，從政者應注意。(鄭宏泰) ▪ 極高和極低的評分也是重要的統計資料。(關焯照等) ▪ 0 分與 100 分都有清楚而具體的定義，不應剔除。若真要剔除 0 分，亦應同時剔除 100 分。即使易除了，平均值仍是低於 50 分。(練乙錚)
關於 50 分是否合格分數	▪ 以 50 分為合格分數並不全面。給予合格分數的人數是佔總受訪人數的 61.8%，給予不合格分數的人數佔總受訪人數的 38.2%。這樣的結果與多年來兩大民研／民調機構定期公布以平均分均多數低於 50 分所顯示的民情相去甚遠。(陳莊勤) ▪ 港大民調的 50 分沒有正面意思，不能視為合格。(鄭宏泰) ▪ 有愈六成人給了 50 分以上的分數。過往新聞報導都把 50 分演繹為合格分數，令市民累積了「50 分等同合格」的印象，港大有必要澄清。(港人講地)	▪ 50 分只是一個中性的分數，沒有所謂合格不合格。(鍾庭耀) ▪ 55 分的支持度大約等如假想投票中的 45%的「得票率」，50 分的支持度則可化成大約 30%的「得票率」，45 分大概會轉化為 20%，而 40 分大概會化成 10%至 15%左右。(鍾庭耀) ▪ 根據問題的措辭，50 分只是代表「一半半」，沒有任何暗示這是一個合格的最低門檻。50 分是評分的中間點，如果把 50 分歸入合格，會得出 61.8%的人給了合格分數。但如果把 50 分歸入不合

	▪ 翻查以往報道，發現鍾庭耀曾多次提到 50 分是合格水平。普通人難以理解 50 分是否合格水平，認為鍾要澄清。(張志剛)	格，會得出 66.4%的人給了不合格分數，兩者結果相反。解決方法是把一半評 50 分的人歸入 0-50 分一組，另一半歸入 50-100 分一組，結果是有 52.4%的人給了 0-50 分，反映特首的支持度評分不是太理想。(關焯照等) ▪ 特首民望評分由 0 至 100，50 分居中，應尊重給予 50 分者的中立態度，不應擅自將「50 分」定義為「合格」。(練乙錚)
數據格式問題	▪ 港大以 SPSS 格式發佈原始數據，必須裝有特定軟件才能開啟。(港人講地)	▪ 開啟 SPSS 格式檔案的軟件，在大學可以免費下載，相信任何一個專業研究機構都有相關軟件。(鍾庭耀)

關於民調的統計學基礎

民調在外國稱為 opinion poll，其要旨是運用統計學方法，找出一個群體對於某個社會議題的意見。統計過程可以分為五大步驟：收集、組織、分析、演繹、發表（《What Is Statistics? – Overview 》）。

做民調的最理想方法是從整個群體（稱為「母體 (population)」）中收集數據，即要訪問群體內的所有人，如此即能得出全面的統計數據，這種做法稱為「人口普查 (population census)」。但現實中往往由於目標群體的人數眾多，只能從受訪對象之中作隨機抽樣 (random sampling) 並進行訪問，這種做法稱為「抽樣統計 (sample statistics)」。無論是人口普查或抽樣統計，在得到原始數據之後，研究員都會組織並分析原始數據以進行總結。最常見的總結方法是取平均值 (mean) 和標準差 (standard deviation)，以展示數據的中央趨勢 (central tendency) 和分散程度 (variability)。中央趨勢的量度，還可以用中位數 (median) 和眾數 (mode)。分散程度的量度還可以用數值範圍 (range，即最大數減最細數)、方差 (variance，即標準差的平方）、百分位數 (percentile) 等。除了中央趨勢和分散程度，有時還要量度數值分布的偏度 (skewness，即非對稱性) 和峰度 (kurtosis，即尖峰的尖銳程度)。這些都是嘗試用少量的數字，去總結一大堆數據的整體特性。數字之外，有時也會用圖表表示數據的特性，最常見的是以直方圖 (histogram) 來展現數據的頻率分布 (frequency distribution)。從上文可知，數字簡潔易用但流於片面，圖表表達較麻煩卻能給出更多方面的資料，研究員在報告中往往要兩者配合使用，才能展現數據的真實特性。

用這些統計結果來描述原始數據的特性，稱為描述性統計 (descriptive statistics)。如果是從樣本的特性來推論整個母體的特性，則稱為推論性統計 (inference statistics) 。中央極限定理 (central limit theorem) 表明，如果樣本數足夠大，而且抽樣足夠隨機，則樣本的平均值會呈正態分布 (normal distribution) 並趨近母體的平均值，而標準差則為母體的標準差除以樣本數的開方。只要符合中央極限定理的條件，便可以從樣本的平均值和標準差，推測母體的平均值和標準差，並推測這些推測的置信區間 (confidence interval) ，以估計可能的誤差範圍，從而決定推測的可信性。然後，研究員便會就著有關調查的主題，演繹並發表調查結果。

關於上述的統計學理論，可以參考一般的統計學入門書籍（如《OpenIntro Statistics》)。

抽樣調查可能出現以下幾種誤差：
其一、因為樣本缺乏代表性而引入誤差。抽樣必然要忽略母體中部份人士的意見，樣本越小，遺漏越多，因此樣本必須要有代表性，即其成份跟母體相若，否則從樣本的特性來推論整個母體的特性時，便會出現誤差 (Wilks, 1940)。例如，有文獻指出部份在美國進行的電話調查，只對家用電話號碼進行抽樣，但現今越來越多人只用手提電話，作者認為有證據顯示這些只用手提電話的人有相當不同的政見，因此以家用電話受訪的樣本不能代表他們 (Mokrzycki, 2010)。

其二、受訪者未必願意表達自己的真實看法。例如問題較敏感，令受訪者不想或不敢表達意見。有學者提出沉默的螺旋 (spiral of silence) 的概念，指出如果受訪者認為自己的意見屬於少數派，便可能不敢發表真實的意見 (Noelle-neumann, 1974)。一項以台灣與美國人為對象的研究指出，接受電話訪問時台灣人展現了沉默的螺旋現象，美國人則不然，顯示某種文化特質可能會導致這現象出現 (Huang, 2005)。

其三、訪問的用語或會影響結果。不同文化、不同背景的人對問題可能有不同的理解 (Groves, 2009)，影響數據的有效性 (validity)。

其四、在總結報告時，無可避免要忽略原始數據中的一些資料。例如平均值的計算方法是將數據總和除以個數，從平均值卻不能反過來計算出原始數據。以 {0, 60, 60} 和 {40, 40, 40} 兩組數據為例，平均值都是 40。兩組數據明顯不同，卻無法從 40 這個數字得知有甚麼不同，因為原始數據的細節被忽略了。如果統計量的選取不宜，便會在演繹出誤導的結果。部份輿論針對平均值所提出的質疑，即屬這一類。

港大民調使用的方法

港大民研網站詳列了特首梁振英評分的相關研究方法(《特首梁振英評分》)。調查基本上每兩個月進行一次，以電話訪問 18 歲以上操粵語的香港市民。每次樣本數為 1000 或以上，抽樣方法是從住宅電話簿中首先以隨機方法抽取「種籽」號碼，在號碼上加減 1 或 2，過濾重覆號碼後再作隨機排列，然後提供給訪員進行電話訪問。如果被抽中的家庭中成員不止一人，就選擇下一位即將生日的家庭成員作訪問。

調查的結果經過了加權 (weighting) 處理。根據上文所引文獻 (Wilks, 1940)，樣本的成份要跟母體相若才有代表性。由於事實並不符合這項要求(例如年齡分布不同)，因此研究員按 2013 的中期人口統計中的性別與年齡分布，及 2011 年人口普查中的學歷分布，對樣本進行了加權，其百分比已詳列於《被訪者基本個人資料》網頁。例如，18-29 歲的人口比例，在原始樣本中為 15.9%，在加權樣本中修正為 18.3%。要留意加權是加在人數上，而不是加在分數上。兩者的概念大有不同。例如一個給了 50 分的人，若要將其所佔的權重加倍，會變成兩個給了 50 分的人，而不是一個給了 100 分的人。有些網站忽略了這一點，錯誤計算出大於 100 分的評分(如:輔仁網《港大民研特首評分係「被拉高」還是「拉低」?》)。調查所用的問卷有幾個版本，關於特首民望的問卷編號為 tp1403013_01 (《調查問卷》)。除了詢問受訪者對特首的支持度之外，問卷還會詢問受訪者的居住地區、家庭成員人數、是否登記選民、有否在各項選舉中投過票、性別、年齡、教育程度、居住情況、婚姻狀況、職業收入、階層(如中產、基層等)、出生地、行業、來港年期等等。

關於特首支持度的問題有兩條：

- Q1: 而家想請你用 0 至 100 分評價你對特首梁振英既支持程度，0 分代表絕對唔支持，100 分代表絕對支持，50 分代表一半半，你會俾幾多分特首梁振英呢？
- Q2: 假設明天選舉特首，而你又有權投票，你會唔會選梁振英做特首？

備受爭議的民望評分即來自 Q1 的答案。基於近日公眾的關注，港大民研網站公布了最近一次(2014 年 3 月 3 日-6 日)的原始數據，檔案格式為 SPSS，內裡包含了 Q1 的數據共 1017 條，亦即此次調查的樣本數。根據 SPSS 檔內的說明，其數據結構如下：

- 第一列：1-1017 的編號；
- 第二列：受訪者所給的 Q1 的分數；其中 3 條記錄是 191，代表「不認識梁振英」。16 條記錄是 8888，代表「不知道」或「不肯講」。餘下 998 條為 0-100 間的整數，即為受訪者給予梁振英的評分。

- 第三列：性別；其中 1 代表男，2 代表女。
- 第四列：年齡組別；其中 1 代表 18-29，2 代表 30-39，3 代表 40-49，4 代表 50-59，5 代表 60-69，6 代表 70 或以上。另有 4 筆記錄是-99，代表拒答。
- 第五列：一個代表權重的數字；例如第一筆記錄的人的權重是 0.85422675557，表示他在經加權處理的樣本中，只代表 0.85422675557 個人。

就著 Q1 的答案，港大民研原先發表的報告中只報告了以下數點（《港大民研發放特首及問責司局長民望數字》）：

1. 特首梁振英的最新支持度評分為 47.5 分，跟兩星期前變化不大。
2. 樣本數是 1017。
3. 回應率是 65.9%。
4. 誤差率是 +/-1.5，即 3% （以 95%置信水平計算）

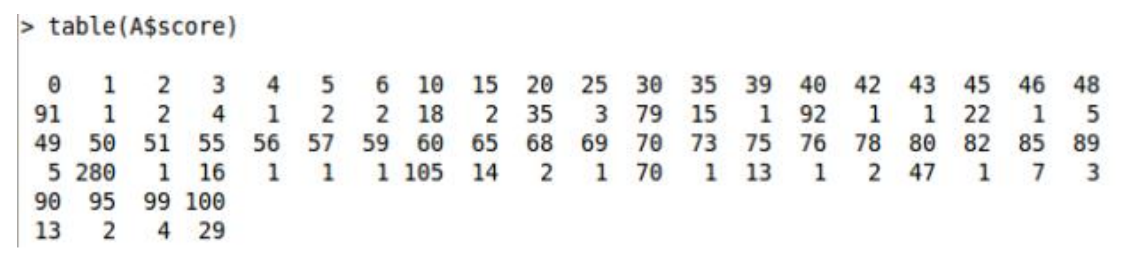
註：報告亦提及，根據民研計劃的標準，梁振英屬於「表現失敗」，其定義為反對率超過 50%。但反對率來自 Q2 的答案，不在本文討論範圍內。有論者認為「表現失敗」是因為梁的平均分在 50 分以下，從而引發關於定義合格分數的批評。按照調查中所用的「民望級別總表」中的定義，這項批評並不符合事實。

分析及評論

參照前述抽樣調查可能出現的幾種誤差，比較港大民研網站所列的研究方法、數據和分析，我們可以評價港大民調在特首民望評分上面的合理與否。
港大民調以電話進行隨機訪問，對種籽電話號碼進行加減處理，並以生日日期選取家庭成員作訪問。最終成功訪問的樣本數達 1000 以上，回應率 65.9%，又對數據進行加權處理，應能很大程度上確保了樣本的代表性。以家用電話號碼來抽樣，可能會出現美國研究中描述的偏頗情況。但目前沒有證據顯示，忽略手提電話的使用者會對關於特首民望的調查造成偏頗的結果，因此不能以此作為對港大民調的指控。

文獻指出人們可能會因為自己的意見屬於少數派而不敢發表真實的意見，即「沉默的螺旋」現象。但是次電話訪問以匿名進行， 應能減低人們的擔憂。而且即使「沉默的螺旋」存在，除非人們認為大多數人都很極端，否則「沉默的螺旋」亦只會令人們傾向選取中間的答案，不會反過來導至「極端答案」的出現。
訪問用語方面，問卷的說明是 0 分代表絕對不支持，100 分代表絕對支持，50 分代表一半半。如果受訪者要從這三個分數中選擇，大部分都會選中間的 50 分。如果要給其它分數，受訪者就要思考其它的數字。圖一顯示各分數的出現頻率，圖二將這頻率以圖象方式表達。從這些數據可知，受訪者傾向給出簡單的數，其中 0 字尾的數字最多（如 0,10,20,30,...），5 字尾的數字較少，其它數字最多只有

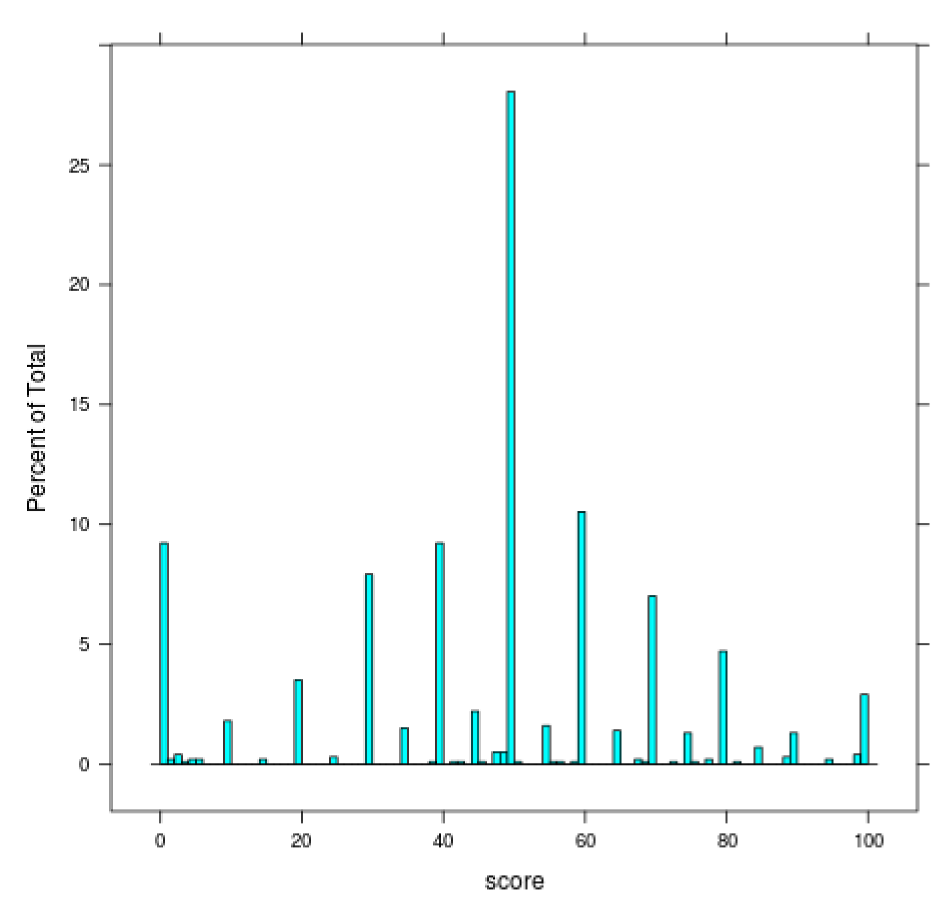
幾個人選擇。另外，選 50 分的人非常多，共 280 人，選 0 分的有 91 人，選 100 分的也有 29 人。這三個分數的出現頻率比旁邊的分數多出很多。理論上，1 分甚或 10 分的相差應該算是輕微的變化，但對受訪者來說，這 0,50,100 三個分數都具有獨特意義。1 分跟 2 分之間可能沒有差別，0 分與 1 分之間的差別卻是巨大的，是質變而非量變。同理，100 分與 99 分之間，49、50、51 分之間的差別亦然。民調要求受訪者給出 0-100 之間的分數，並以此計算平均值，是假定了這個分數跟受訪者心目中對特首的支持度之間有一連續變化的線性關係。事實上，問題的問法賦予了三個分數特別的意思，客觀上扭曲了分數分布。這效應在 50 分這一臨界點尤為重要，下面再詳述。



```
> table(A$score)

  0   1   2   3   4   5   6  10  15  20  25  30  35  39  40  42  43  45  46  48
 91   1   2   4   1   2   2  18   2  35   3  79  15   1  92   1   1  22   1   5
 49  50  51  55  56  57  59  60  65  68  69  70  73  75  76  78  80  82  85  89
  5 280   1  16   1   1   1 105  14   2   1  70   1  13   1   2  47   1   7   3
 90  95  99 100
 13   2   4  29
```

圖一：各分數的頻率分布

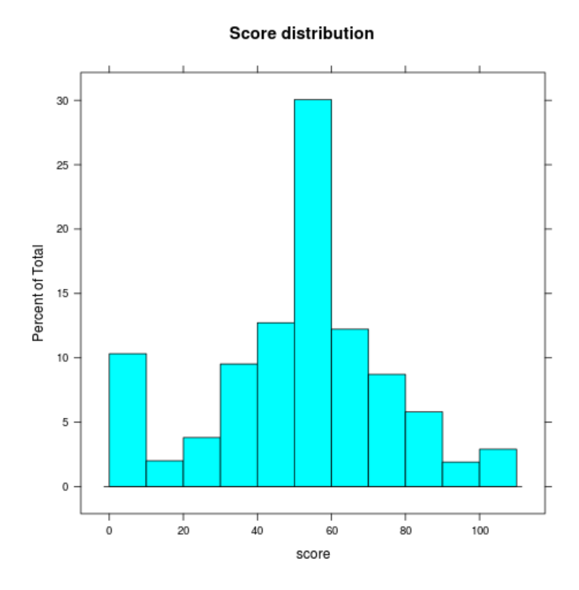


圖二：分數的頻率分布圖（以 1 分為一格）

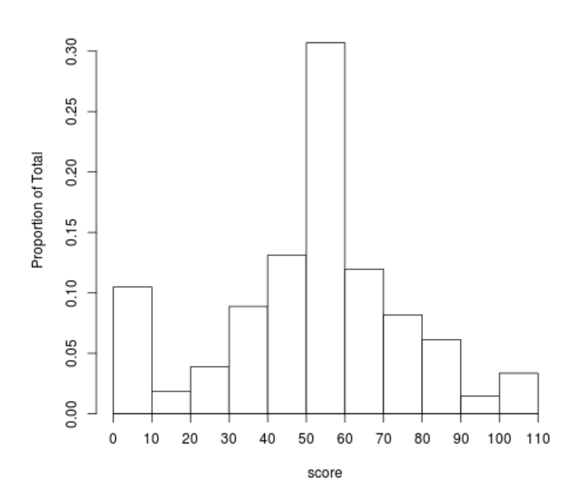
原報告以報導平均分為主，新聞媒體主要亦以這個數字作為討論的根據。如前所言，平均分只是總結統計數據的其中一種方式，不同的統計量會給出不同方面的資訊。平均分是最常用的方式，其好處是計算涉及所有的數據，壞處是易受極端數字影響。如果數據中出現極端的數字，一般做法是以中位數取代。中位數是指將數據順序排列之後排在中間的數。例如，數集 {0,0,0,0,100} 的平均值是 20，中位數是 0。平均值因受 100 影響，其數值不能很好地反映數集的中央趨勢。反之，中位數只取決於數字的排列，在這情況下就較能反映中央趨勢，這就是為甚麼入息通常都是以中位數而非平均值來計算中央趨勢。至於眾數，則是頻率最高的數，在這例子也是 0。也有一些情況是三個數字都不能很好地反映中央趨勢。例如，數集 {0,0,0,100,100,100} 的平均值是 50，中位數是 50（中間兩個數的平均），眾數是 0 和 100 （因頻率相同），三個數字都難以代表數集的總體特性，因為數集本身就是分化成兩邊的。一般來說，只有當分布接近鐘形分布時，這三個統計量才能較好地反映現實。

從原始數據可知，是次民調的分數分布並不依從鐘形分布，單純從數字很難對統計結果作出全面的認識，因此以下改由圖表進行分析。

圖三是以每 10 分為一組的頻率分布，是未經加權處理的結果，分組方法為 0-<10、10-<20、20-<30、30-<40、40-<50、50-<60、60-<70、70-<80、80-<90、90-<100，100-<110。留意最後一個分組實際上只有 100 分的分數。一般做法是把 100 分歸入前一組，變成 90-100。但因在這組數據中，100 分出現了峰值，所以做了這個特別處理，以免影響了前一組的結果。加權處理則按各權重調整每一組的頻率，分組方法相同，結果如圖四所示。



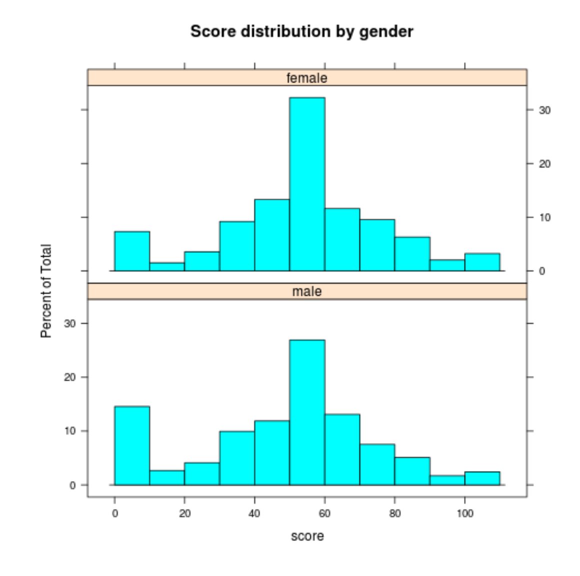
圖三：未經加權處理的頻率分布



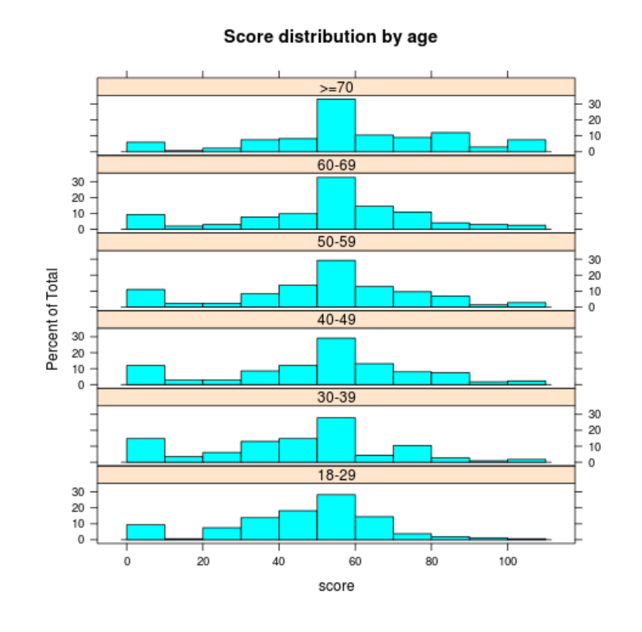
圖四：經過加權處理的頻率分布

兩幅圖只有些微差別。由於本文的分析以看圖表為主，不涉及計算合格不合格的問題，為了方便說明，以下將採用未經加權處理的頻率分布。

跟圖二的結果一樣，圖三清楚展現了 0 分、50 分和 100 分的特殊性。除了總體的分布外，港大公佈的原始數據還包括年齡和性別的資料，因此我們也可以按性別和年齡分別畫出各組別的分布，如下面兩幅圖所示。



圖五：以性別分組的分數分布



圖六：以年齡分組的分數分布

先看 0 分的情況。無論是按性別還是年齡分組，都可以看到 0-<10 分處出現尖峰。從原始數據或圖二都可以看出，在這個組別裡絕大部分都是直接給了 0 分。進一步說，男性受訪者給 0 分的人較女性多，有接近 15%。而 30-39 歲的組別給 0 分的人較其它組別多，亦是接近 15%。從 40 歲開始，年紀越大的組別，越少人給 0 分。即使忽略了這些給 0 分的情況，也可以看出 18-29 歲及 30-39 歲的市民，評分少於 50 分的較評分多於 50 分的為多。而隨著年紀增加，排除 0 分之後兩邊趨向平衡。到了 60-69 歲及 70 歲或以上的組別，則有向右邊發展之勢。因此，如果以給 0 分的作為對特首極度不滿的標示，則可以看出最不滿特首的是介乎 30-39 歲的市民。從 40 歲的組別開始，年紀越大的市民對特首的支持度越高。18-29 歲是剛剛畢業出來工作的年紀，30-39 歲是成家立業的年紀。這兩個年齡層的不滿，或許反映了政府在經濟、就業等政策上的不足，也有可能是這個年齡層的人較關心政治，尤其是在民主發展上產生不滿。真正原因必須經進一步研究確定，本文只能從數據上指出這一現象，沒有足夠的資料作出解釋。

再看 50 分和 100 分的尖峰。明顯的 100 分尖峰只出現在 70 歲或以上的組別。事實上，70 歲或以上的組別，50 分尖峰兩邊的分布很均勻，而 50 分尖峰比其它組別都突出。圖二的分布也顯示，50 分尖峰的人數，遠遠超出了鐘形分布應有的數量。透過比較旁邊兩組的高度，大約也是多了 15%。如前所述，問題的設計很容易令人選擇 50 分。這些人要麼真是覺得自己對特首的支持度是一半半，也有可能只是覺得難以下決定，或者根本沒有打算認真思考這個問題，只好給一個中間的分數。如果這班人經過了詳細思考，就可能會給出較高或較低的分數。鑑於這班人的人數不少，他們的決定會對整體分布產生關鍵影響。無奈問卷的設計無法把這批人分辨出來，因此我們不知道這班人的真正取態。

總結及建議

本文透過統計學分析方法，嘗試檢視近日輿論對港大民調的批評及反駁，探討這些言論背後的統計學理據。本文作者認為，港大民調在抽樣方面十分嚴謹，但在設計問卷和演繹結果方面有值得適榷之處。

其中，無論以平均分、中位數還是眾數來進行統計，都不能全面地反映調查結果。應該同時公布頻率分布，甚至是各年齡組別的頻率分布，才能從中提出改善施政的建議。在分析極端分數的時候，我們可以把這些分數分開來考慮，以反映其他人的意見，但極端分數還是有它的重要價值。至於給予 50 分的人數眾多，本文認為是來源於問卷設計出現了問題，致使難以得知這些人的真正取態，降低了調查的價值。

關於合格分數的問題，由於原問卷設計中，50 分只是一半半的意思。以 50 分為合格分數可能符合一些人的直覺，但本文認為沒有壓倒性的理由以此定義為合格分數。合格是最低要求的指標，但這個最低要求設在何處則是沒有一定準則。即使在學校的考試制度裡，合格分數也並非每間學校相同，只能說通常在 40-60 分之間。本文同意鍾氏的說法，50 分只是一個中性的分數，沒有必要跟合格不合格掛鈎。傳媒亦不應再以此作為報導的焦點。

此外，從按年齡組別畫出的分數分布可以看出，民調的數據確能反映一些重要的社會現象 。雖然大多數人中間落墨，所謂的極端分數只佔少數，但亦有一成之眾，而且集中在 30-39 歲的組別。在一個社會裡，沉默的大多數和激進的極少數同樣重要。前者是社會穩定的要素，後者是變革的動力，缺一不可。為甚麼某些組別的人給了最差的評分，他們最關注的是甚麼，這方面的跟進工作，不但能夠回應這組人的關注，亦有可能帶動社會的整體進步，從政者責無旁貸。

最後，本文作者很感謝港大民研公開最近一次民調的原始數據，讓社會大眾可以進行更深入的分析。然而 SPSS 只是學術界常用的統計軟件，但如果數據的使用對象是傳媒或一般大眾，通常的做法是一併提供 CSV 和 Excel 版本，有時也會提供 XML 版本（參看：美國政府的《Data.gov》、香港政府的《資料一線通》）。現時在 MS Excel 上開啟 SPSS 格式檔案必須另外安裝插件，本文作者亦是使用了 PSPP （《PSPP – GNU Project – Free Software Foundation》）或在 R （《The R Project for Statistical Computing》）安裝某些特定的程序包才能開啟。若能以比較普及的格式提供數據，將有助資訊的透明和公開。

其他參考資料

華生：〔中國〕城鄉差距的統計誤導和真實挑戰
媒体称统计造假误导决策 病根在于数字出官
人民日报批评地方 GDP 报花账：误导宏观决策
但愿失真的统计数据不会误导个税决策
林美芬：內地數據「灌水」 誤導中央害經濟

騙人的誠實數字：談談《歐盟動物園報告 2011》於對比圈養及野生海豚死亡率數據時所做的誤導

潘震澤：民調可靠嗎？

蕭亮思：為甚麼「小學雞統計」應納入通識？